Análise de redes sociais em dados bibliográficos

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Urubatan Rocha Pacheco e aprovada pela Banca Examinadora.

Campinas, 8 de Novembro de 2010.

Ricardo de Oliveira Anido (Orientador)

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecária: Maria Fabiana Bezerra Müller – CRB8 / 6162

Pacheco, Urubatan Rocha
P115a Análise de redes sociais em dados bibliográficos/Urubatan Rocha Pacheco-- Campinas, [S.P. : s.n.], 2010.

Orientador : Ricardo de Oliveira Anido.
Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Computação.

1.Redes de relações sociais - Modelos matemáticos. 2.Algoritmos. 3.Análise por conglomerados. 4.Recuperação da informação. 5.Classificações bibliográficas. 6. Indexação automática. 7.Ciência da informação. 8.Banco de dados bibliográficos. I. Anido, Ricardo de Oliveira. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Título em inglês: Social network analysis on bibliographical data

Palavras-chave em inglês (Keywords): 1. Networks, Social – Mathematics models.
2. Algorithms. 3. Cluster analysis. 4. Information retrieval. 5. Bibliographic classification.
6. Automatic indexing. 7. Information science. 8. Data libraries.

Área de concentração: Metodologia e Técnicas da Computação

Titulação: Mestre em Ciência da Computação

Banca examinadora: Prof. Dr. Ricardo de Oliveira Anido (IC – UNICAMP)
Prof. Dr. Nabor das Chagas Mendonça (CCT – UNIFOR)
Prof. Dr. Reinaldo Alvarenga Bergamaschi (IC - UNICAMP)

Data da defesa: 08/11/2010

Programa de Pós-Graduação: Mestrado em Ciência da Computação

TERMO DE APROVAÇÃO

Dissertação Defendida e Aprovada em 08 de novembro de 2010, pela Banca examinadora composta pelos Professores Doutores:

Prof. Dr. Nabor das Chagas Mendonça
CCT / UNIFOR

Prof. Dr. Reinaldo Alvarenga Bergamaschi
IC / UNICAMP

Prof. Dr. Ricardo de Oliveira Anido
IC / UNICAMP

Instituto de Computação
Universidade Estadual de Campinas

Análise de redes sociais em dados bibliográficos

Urubatan Rocha Pacheco[1]

Novembro de 2010

Banca Examinadora:

- Ricardo de Oliveira Anido (Orientador)
- Nabor das Chagas Mendonça - CCT/UNIFOR
- Reinaldo Alvarenga Bergamaschi - IC/UNICAMP
- Jacques Wainer - IC/UNICAMP (Suplente)
- Roberto M. Cesar Jr. - DCC do IME/USP (Suplente externo)

[1]Suporte financeiro de: Bolsa da Fapesp (processo 08/54276-9) 2008–2009

Resumo

O foco deste trabalho é viabilizar a análise estrutural em redes sociais de colaboração científica a partir de bases de dados bibliográficos. Os dados bibliográficos são utilizados para obter redes sociais de afiliação dos autores a instituições de pesquisa científica, e das publicações são extraídas as suas relações com ontologias de áreas de pesquisa.

Foram estudados e aplicados métodos que utilizam a análise das redes sociais para solução/redução de ambiguidades em identidades de nomes de pesquisadores, instituições, e veículos científicos. Outro assunto estudado foi a abordagem de medida da qualidade dos resultados e os problemas que afetam a sua qualidade.

Concretizando o objetivo deste trabalho, foram construídas métricas e ferramentas que permitem a comparação da produção científica entre instituições, departamentos, áreas de pesquisa, países, etc. As ferramentas também produziram um *ranking* de universidades baseado no prestígio dos pesquisadores destas universidades na rede social de coautoria. Este resultado permitiu demonstrar que a informação estrutural de prestígio foi devidamente capturada ao correlacionar este *ranking* com outros que avaliam a qualidade da produção científica das universidades utilizando critérios semelhantes.

Abstract

This work performs social network analysis of the scientific collaborations extracted from bibliographic data bases. The analysis also includes the authors' scientific institution affiliation, and its relation with the main scientific publications and with research subject ontologies.

We studied and applied methods that use social network analysis to solve or mitigate the problem of ambiguity in researchers' identities. We also applied the methods for ambiguity resolution for names of institutions, scientific meeting venues, country/state names, etc. Another study subject was measuring the quality of the results.

Finally we developed metrics and implemented tools that allow the comparison of the scientific production of institutions, researcher groups, research subjects fields, countries, etc. The tools also produced a ranking of universities based on the prestige of these universities researchers at the co-authorship social network. These results demonstrated that prestige structural information was properly captured showing its correlation with other works that assess the quality of scientific production of universities using similar criteria.

Agradecimentos

Primeiramente gostaria de agradecer a minha esposa por me apoiar em todas as etapas do mestrado, me permitindo concentrar integralmente no trabalho e por diversas vezes ter alterando seu horário e aumentando ainda mais o seu encargo nos afazeres domésticos.

Agradeço também ao professor Ricardo Anido, por estar sempre disponível ao diálogo, por sempre estar disposto a ouvir, contribuir com ideias e direcionamentos para a boa qualidade do trabalho. Desta forma, me orientou neste trabalho de Mestrado e me orienta no de Doutorado.

Agradeço à FAPESP pelo apoio financeiro, pois sem ele não seria possível desenvolver este trabalho.

Aos colegas, amigos, familiares e professores que contribuíram dialogando e trocando ideias e compartilhando ferramentas, ofereço também o agradecimento por desta forma terem contribuíram para o trabalho.

E por fim agradeço a minha mãe e postumamente a meu pai, que carinhosamente sempre me incentivaram a investir no meu desenvolvimento acadêmico. Meu pai ficou orgulhoso em saber que consegui criar condições de iniciar o Mestrado, mas infelizmente faleceu em seguida.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

O objetivo deste trabalho é o desenvolvimento de ferramentas e de uma metodologia para análise de redes sociais em dados bibliográficos de um dado grupo de autores, com vistas à comparação quantitativa e qualitativa com outros grupos de autores. Uma das aplicações das ferramentas desenvolvidas é a criação de *rankings* de departamentos de Universidades, possibilitando a comparação entre departamentos e/ou entre áreas de conhecimento.

Dada uma base de informações bibliográficas de artigos publicados em diversos veículos (autores, ano de publicação, categorias de áreas de pesquisa, etc.) e uma lista contendo os nomes de autores, foi construída uma rede social baseada em dados desta base (coautoria, afiliação, áreas de conhecimento, etc.). O conjunto de ferramentas que foi construído é capaz de extrair medidas de análise de redes sociais dos dados bibliográficos. Esta ferramenta pode ser utilizada, por exemplo, para comparar grupos de autores e analisar a produção científica de departamentos e universidades em qualquer área de pesquisa. Os principais problemas que foram tratados são:

- a redução de ambiguidade na denominação de entidades (autores, afiliação, veículos e outros atributos que possam ter variações desconhecidas nos seus nomes) [27],
- metodologias de agrupamento destas várias denominações que representam uma determinada entidade e
- funções de similaridade entre nomes destas entidades [37].

O presente trabalho aborda problemas que não foram totalmente resolvidos em sistemas importantes de acesso à base de dados bibliográficas como WOS [39] e Scopus [7] com relação à identificação única dos autores dos artigos. Para tratar este problema foi utilizada a análise da estrutura das redes sociais para solução/redução de ambiguidades em identidades de uma base de dados bibliográficos. Desta forma, se produziu um mecanismo mais preciso de sumarização de informações nos dados bibliográficos.

Para demonstrar estes conceitos foi utilizado como fonte de dados bibliográficos um banco de dados extraído da WOS ISI [39] das universidades de maior renome dentre as brasileiras e estrangeiras. Os dados se referem aos artigos produzidos de 2003 a 2007, e foram extraídos no início de 2008. O grupo de autores escolhido foi o quadro de professores dos departamentos de computação e engenharia da computação destas universidades, com vistas à comparação quantitativa e qualitativa entre estes grupos.

Através dos experimentos, foi possível concluir que apesar de um mecanismo de agrupamento simples, e de funções de similaridade com definições de parâmetros apenas intuitivos, foi possível um bom resultado ao recuperar as informações exatamente por capturarem informações pertinentes dos relacionamentos entre as entidades de dados envolvidas. Foi possível mostrar também que a estrutura da rede de coautoria pode ajudar a melhorar os resultados, principalmente no aspecto da precisão em detrimento da medida de recuperação. Os indícios de similaridade tiveram um papel essencial no sucesso desta abordagem. Estes indícios de similaridade são as informações sem relação unívoca direta com as denominações dos autores, mas apenas ao artigo (lista de emails de autores do artigo, lista de afiliações dos autores do artigo, lista de categorias de pesquisa do artigo, nome do veículo).

Uma rede de coautoria foi produzida pela ferramenta. E esta foi utilizada para produzir um *ranking* de universidades baseado no prestígio dos pesquisadores destas universidades na rede social de coautoria. O coeficiente de correlação de *rankings Kendall* [41] obteve um valor significativo quando comparando aos *rankings* que focam em áreas de pesquisa como o ARWU [12] para campo área de pesquisa engenharia/tecnologia e ciência da computação dos anos 2007 e 2008, para área de ciência da computação de 2009 e com o *ranking* USNews [31] para área de ciência da computação de 2010. Todos estes *rankings* medem a reputação dos departamentos e institutos da computação e engenharia da computação. Isto demonstra que o *ranking* baseado em status na rede de coautoria captura uma medida estrutural das relações de prestígio e que ainda se mantém no decorrer dos anos, apesar das variações que ocorreram de ano para ano em cada *ranking*. Ou seja, a rede de coautoria foi capaz de capturar uma informação de prestígio semelhante a estes *rankings*.

A expectativa é que a mesma metodologia como um todo possa ser aplicada para outras áreas de pesquisa.

Este documento está organizado da seguinte forma: no capítulo 2 são apresentados os principais conceitos que envolvem a metodologia proposta, em conjunto com a revisão bibliográfica de trabalhos relacionados com este projeto. A seguir o capítulo 3 descreve a ferramenta que implementa a proposta deste trabalho; os resultados da aplicação de métodos que utilizam a análise de redes sociais estão no capítulo no capítulo 4, onde também são discutidos alguns aspectos da qualidade dos resultados. Finalmente, as conclusões e trabalhos futuros estão descritos no capítulo 5.

Capítulo 2

Conceitos e revisão bibliográfica

Alguns dos principais conceitos necessários para formar a base do projeto são vistos nas próximas subseções. Neste capítulo também é encontrada a revisão bibliográfica dos trabalhos relacionados.

2.1 Análises de Redes Sociais

A definição de rede social vem da área de ciências sociais:

- "as ciências sociais focam na estrutura dos grupos humanos, comunidades, organizações, mercados, sociedade, ou o sistema mundial. A estrutura social pode ser contextualizada como uma rede de laços sociais" [14];
- "Dada uma coleção de atores, a Análise de Redes Sociais pode ser usada para estudar as variáveis estruturais medidas neste conjunto de atores. A estrutura relacional de um grupo, ou um grande sistema social, consiste num padrão de relacionamento entre coletivos de atores" [45];
- "O Analista de redes sociais deve buscar modelar estes relacionamentos para desvendar a estrutura de um grupo. Ele pode, então, estudar o impacto desta estrutura nos indivíduos dentro do grupo, e ou a influência da estrutura no funcionamento do grupo." [45].

Neste trabalho são utilizados dados bibliográficos como fonte da informação para construir e analisar algumas de suas redes sociais, e assim, mais adiante no texto, são apresentadas algumas métricas de análise de redes sociais que são úteis para o trabalho. Alguns sistemas já realizam cálculos de métricas de avaliação de produção científicas e nos serviram como exemplo, dos quais se destacam o *web of science* [39], e *scimago jcr* [19].

Definições básicas em redes sociais

A análise de redes sociais se baseia na premissa de que os relacionamentos entre os atores sociais envolvidos possam ser descritos por um grafo. Os vértices do grafo representam os atores sociais. Os arcos deste grafo conectam os pares de vértices e sendo assim, estes representam interações sociais. Esta representação permite que pesquisadores apliquem a Teoria dos Grafos [45] para a análise de problemas que de outra forma poderiam ser considerados intrinsecamente vagos e de passíveis apenas de compreensão superficial: este "emaranhado das nossas relações sociais".

Assumindo este tipo de representação mais precisa, uma rede social pode ser vista como um grafo $G = (V, E)$ no qual os vértices V representam os atores sociais, e os arcos E representam os laços de interação social entre os atores. Este grafo da rede social pode ser descrito em termos de suas propriedades em dois níveis: métricas globais do grafo e propriedades individuais dos atores. As métricas globais do grafo procuram descrever as características da rede social como um todo, por exemplo: o diâmetro do grafo, a distância média entre os vértices, o número de componentes (subgrafos conexos maximais do grafo), o número de cliques, etc. As propriedades de atores se referem à análise das propriedades individuais dos atores da rede, por exemplo: o seu status social, a sua distância e posição dentro de um grupo.

Clustering (Agrupamento) descreve a situação em que se um ator qualquer tem relação com dois outros atores, então, há uma razoável probabilidade de que estes dois indivíduos terão algum contato também. Isto pode ser percebido, por exemplo, observando dois amigos e os amigos que cada uma destas pessoas possui. Segundo Watts é bem provável que haja certa intersecção entre estes dois conjuntos de pessoas [46]. O coeficiente de agrupamento é a probabilidade de que dois nós associados a um nó sejam associados eles mesmos entre si no grafo que representa a rede social. Um alto coeficiente de agrupamento indica uma alta coesão (*cliquishness*). Se o coeficiente de agrupamento é significativo e as características do tamanho do caminho na rede sociais satisfazem certas condições, então a rede social observada é de um tipo específico de rede, chamado de *small world network*.

O status de um ator é usualmente expresso em termos de sua centralidade. Por exemplo, a medida do quanto central ele está no grafo da rede. Atores centrais são bem conectados com outros atores e métricas de centralidade irão em geral medir: o seu grau em teoria dos grafos (o número de arcos que incidem nele), a distância média para os outros atores, ou a fração dos caminhos geodésicos que passam por quaisquer outros pares de atores da rede e que passam por este indivíduo.

Uma classe de métricas de reputação objetiva expressar a natureza recursiva do status, ou seja, quando algo tem conferido a sua reputação por um ator de status alto, isto aumenta o status do que foi referido mais do que se fosse destacado por um ator de baixa reputação. Desta forma, o status de um ator pode ser derivado do status dos atores aos

quais ele tem relações que destacam sua reputação. Isto leva a uma definição recursiva de status que matematicamente pode ser definida como análise de autovetores. Considerando que a estrutura dos *hiperlinks* da web mimetiza as propriedades de um grafo de rede social, onde os nós são as páginas, e os *hiperlinks* as arestas, a análise de autovetores pode ser usada para medir o prestígio das páginas da internet; o algoritmo mais conhecido é o *PageRank* [6].

Neste projeto, a expectativa é encontrar redes de coautoria, e em outras relações com as entidades envolvidas, que tenham as características de *small world network*. As *small world networks* são a base para novos algoritmos para medida de similaridade. A principal rede social considerada neste trabalho é a rede de coautoria, construída a partir da base de dados de uma Biblioteca de Dados Bibliográficos Digital. Os atores são os autores listados na base. Uma relação entre os autores é inserida no grafo se os autores colaboraram em pelo menos uma publicação. O uso de redes de coautoria, ou relacionamentos de coautoria, em geral não é novidade na atividade de encontrar sinônimos [5] [22] [20] [21] [28]. Além disso, o objetivo é obter medidas relacionadas a grupos de indivíduos, que de forma análoga poderiam ser consolidados como atores numa rede social que representem a relação entre estes grupos.

Redes Sociais Multimodo e redes de afiliação

Segundo S. Klink *et al.* [24], para se representar informações adicionais de relacionamentos são necessárias as Redes Sociais Multimodo. Segundo Wasserman "o modo em redes sociais se referem a um conjunto distinto de entidades no qual as medidas de redes sociais são feitas" [45]. Assim, em contraste com a rede social uni modal, em que temos autores como atores e coautoria como relações, as redes multimodo são capazes de representar relações de pertinência a conjuntos como: afiliação de professores a suas instituições, artigos ao seu veículo de publicação ou conferências e periódicos, etc. Estes tipos de redes são também conhecidos como *affiliation networks* ou *membership networks* onde um conjunto de atores (neste caso: autores) e vários outros conjuntos de eventos (no caso: publicações, afiliações, conferências, periódicos, etc.) são representados [45]. Relacionamentos contidos em dados bibliográficos podem ser listados como uma hierarquia com indireção crescente:

1. autores contidos numa mesma publicação (ex. coautores),
2. coautores de coautores (ex. amigo-do-amigo),
3. autores na mesma conferência (assuntos em periódicos) (ex. DEXA'06),
4. autores da mesma linha de tendência na conferência (periódico) (ex. VLDB),
5. autores de conferências similares (periódico),

6. autores com publicações similares (p. ex. utilizando palavras chaves),
7. ontologia ou hierarquia de conceitos x veículos de publicações [16].
8. autores do mesmo departamento, instituto, universidade ou empresa [16].

2.2 Métricas em Análise de Redes Sociais

Nesta seção são apresentadas algumas métricas utilizadas em análise de redes sociais.

Proximidade (*Closeness*)
A proximidade mede o quanto um indivíduo está próximo a todos os outros indivíduos da rede (direta ou indiretamente). Ou seja, isto reflete sua habilidade para acessar informações através das "fofocas" (*grapevine*) dos membros da rede. Assim, proximidade (*closeness*) é o inverso da soma de distâncias dos caminhos mais curtos entre cada indivíduo da rede, em relação a todos os outros indivíduos na rede.

Grau/Centralidade (*Centrality*)
A centralidade é a contagem do número de ligações com outros atores na rede. Ou seja, é o Grau de um nó em teoria dos grafos.

Centralidade do Fluxo (*Flow betweenness centrality*)
A centralidade de fluxo é quanto um nó contribui para a soma do fluxo máximo entre todos os pares de nós (exceto este nó). Ou seja, para cada nó são contabilizados quantos caminhos mínimos entre outros dois nós passam por ele, dando assim um indicador de quanto cada nó contribui para o fluxo máximo da rede.

Centralidade de Autovetores (*Eigenvector centrality*)
A centralidade de autovetores é a medida da importância de um nó na rede. Esta medida associa pontuações relativas para todos os nós da rede baseado no princípio que conexões vindas de nós que tem alta pontuação contribui mais para o nó em questão. Semelhante ao *PageRank*.

Centralização (*Centralization*)
A centralização é a diferença entre n das ligações para cada nó nesta rede dividido pela máxima soma possível das diferenças numa rede que tenha o mesmo número de vértices e seja um grafo completo. Uma rede centralizada terá muitas de suas ligações dispersas em torno de um ou poucos nós, enquanto uma rede descentralizada terá pouca variação entre as n ligações possuídas por cada nó.

Coeficiente de agrupamento (*Clustering coefficient*)

O coeficiente de agrupamento é a probabilidade de que dois nós associados a um nó sejam associados eles mesmos. Um alto coeficiente de agrupamento indica uma alta coesão (*"cliquishness"*) [14].

Coesão (*Cohesion*)

A coesão mede a proporção em que os atores são conectados entre si por laços coesos. Estes grupos identificados como "blocos" (*cliques*) se todos os atores deste grupo (subgrafo) são conectados aos outros atores deste mesmo grupo, "círculos sociais" ('*social circles*') se o contato é menos direto (ou seja, são conjuntos de *cliques* com intersecções entre si) [14].

Densidade (*Density*)

A densidade mede a proporção entre o número de ligações da rede, em relação ao número máximo de ligações possíveis, em uma rede com este número de nós, ou seja, num grafo completo [14].

Comprimento do caminho (*Path Length*)

O comprimento do caminho é a medida da distância entre pares de nós na rede. O comprimento médio de caminho é a média destas distâncias entre todos os nós da rede.

Alcance (*Reach*)

O alcance mede o grau em que qualquer membro da rede pode alcançar outros membros da rede, ou seja, corresponde em teoria dos grafos ao diâmetro do grafo.

Buracos estruturais (*Structural hole*)

Os buracos estruturais são definidos como o conjunto mínimo de membros que caso sejam removidos da rede podem aumentar o número de componentes (subgrafos conexos) nesta rede. Esta medida está relacionada à ideia de capital social: se um indivíduo está relacionado a dois outros que não estão diretamente ligados entre si, o primeiro poderá controlar a comunicação entre os demais.

2.3 Redes de Coautoria

Redes de coautoria constituem uma classe importante de redes sociais, e estas têm sido usadas amplamente para determinar a estrutura das colaborações científicas e o status individual dos pesquisadores. Embora seja um pouco similar às redes de citação em literatura científica, coautoria implica em um laço de relacionamento social muito mais forte do que a citação. As citações podem ocorrer sem que os autores se conheçam uns aos

Tabela 2.1: Artigos utilizados nos exemplos de redes de coautoria

Artigos		Autores
artigo 1	$\rightarrow$	$\{v_1, v_2, v_3\}$
artigo 2	$\rightarrow$	$\{v_1, v_2\}$
artigo 3	$\rightarrow$	$\{v_1\}$

outros, e pode se estender através do tempo. Coautoria implica em relacionamentos entre colegas, e que ocorrem de forma contemporânea aos participantes colocando a análise de redes sociais neste tipo de rede num nível de confiança maior do que as redes de citação.

O exemplo mais antigo de redes de coautoria é o Projeto do Número de Erdös, no qual o menor caminho através dos laços de coautoria entre qualquer matemático e o matemático Húngaro Erdös foi calculado [13]. Newman estudou e comparou o grafo de coautoria do arXiv, Mediline, SPIRES e NCSTRL [35] [34] e encontrou algumas diferenças entre os resultados da análise para disciplinas teóricas e experimentais, como por exemplo, a presença de uma quantidade menor de coautores em artigos de pesquisa teórica, em relação aos artigos de pesquisa experimental.

2.3.1 Construção de Redes de Coautoria

Nesta seção são apresentados aspectos fundamentais da representação destacando a abordagem usada para representar a rede de coautoria. O modelo mais comumente usado em redes de coautoria é um grafo G binário, não orientado, no qual cada arco representa um relacionamento de coautoria. O segundo modelo considera um grafo com relações binárias orientadas que permite o cálculo do prestígio dos atores da rede.

Redes de coautoria binárias e não orientadas

O modelo de redes de coautoria mais simples e mais usado é o baseado em um grafo G no qual cada arco representa um relacionamento de coautoria.

Este modelo pode ser construído da seguinte forma: dado um conjunto de autores de artigos, então, se para algum par de autores ambos forem coautores de um artigo, um arco de peso unitário é criado. Na Figura 2.1 item *a* é apresentado um exemplo com três artigos conforme a Tabela 2.1, onde os autores v_1 e v_2 são conectados por um arco pois são coautores do *artigo 1* e assim por diante. O grafo resultante é denominado grafo não orientado de peso unitário $G = (V, E)$, onde o conjunto de n autores é definido por $V = \{v_1, ..., v_n\}$ e $E \subseteq V^2$ que representa os arcos entre os autores. Como será visto nas seções seguintes, várias métricas de grafos podem ser calculadas para este tipo de rede.

Redes de coautoria binárias e orientadas

Para que seja possível a medida de prestígio de um autor, é necessário distinguir o "aval" concedido do "aval" recebido pelos autores. Na análise de redes sociais, o conceito de prestígio é definido por relações direcionadas. Desta forma, é preciso converter o grafo de coautoria (que é por definição não orientado) em um grafo orientado. Para isso, foram consideradas algumas premissas:

- qualquer rede não direcionada pode ser representada em uma rede direcionada com arcos simétricos, por exemplo, todos os arcos de um grafo G são substituídos por outros dois, simetricamente direcionados.

- estes arcos simétricos e direcionados representam um endosso mutuo entre os autores, ou seja, em um passeio aleatório, os arcos direcionados podem ser entendidos como um movimento possível nas duas direções do percurso.

- os arcos têm peso binário, indicando a presença ou ausência destes dois relacionamentos simétricos.

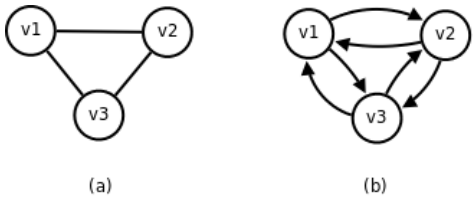

Figura 2.1: Representações de redes de coautoria: (a) grafo binário não orientado; (b) grafo binário orientado.

O grafo resultante deste processo é denominado grafo orientado de peso unitário (veja na Figura 2.1 item *b*). Como será apresentado na seção 2.3.2, o algoritmo de *PageRank* e outras medidas de prestígio podem ser aplicados neste modelo de rede. Uma observação importante é que esta representação poderia ser estendida de forma a atribuir pesos diferentes para o prestígio recebido e concedido, com base na intensidade da atividade de coautoria entre os autores, em relação à intensidade de publicação do autor, que confere prestígio ao outro. No entanto, não foi realizada esta abordagem neste trabalho, pois exigiria uma alteração no algoritmo do *PageRank*. Mas, no artigo sobre o *AuthorRank* proposto por Xiaoming Liu [29], podem ser encontrados mais detalhes sobre esta abordagem. Segundo este artigo, o *AuthorRank* tem um resultado um pouco melhor que o obtido pelo *PageRank* para capturar o prestígio dos autores na rede de coautoria.

2.3.2 Métricas para Redes de Coautoria

Nesta seção são apresentadas as métricas que se aplicam aos modelos de representação de redes de coautoria apresentados na seção anterior. Assim, na análise de redes sociais existem diversas métricas disponíveis para medir as características de uma rede de colaboração na forma de um grafo binário não orientado, incluindo análise de componentes, análise de *small world*, e análise de centralidade. Estas métricas medem várias propriedades e algumas somente podem ser aplicadas em certas condições. As métricas usadas neste trabalho são listadas na Tabela 2.2 destacando sua aplicabilidade. Estas métricas são discutidas mais em detalhes logo a seguir.

Análise do tamanho dos componentes da rede social

Um componente de um grafo é um subconjunto dos vértices deste grafo no qual exista algum caminho de um vértice deste subconjunto e qualquer outro vértice neste subconjunto. Uma rede de coautoria geralmente é formada por vários componentes desconexos (por exemplo, grupos de pesquisa e pessoas que publicam sozinhas), e a análise de componente pode ser usada para o estudo sobre a estrutura da rede. Alguns métodos de análise de rede são na maioria aplicados em redes conexas. No entanto, em redes com componentes desconexos, estes métodos são tipicamente aplicados no maior componente conexo, como mostra a Tabela 2.2.

Tabela 2.2: Métricas em redes de coautoria

Métrica	Propriedade		Escopo		Importância	
	Ator	Global	Rede inteira	O componente maior	Centralidade	Prestígio
Component		x	x			
Small world		x		x		
Closeness	x			x	x	
Betweenness	x		x		x	
Grau	x		x		x	
PageRank	x		x			x

Grau, proximidade e centralidade do fluxo

Nas análises de redes sociais, as três métricas mais comuns são denominadas de "grau, proximidade e centralidade do fluxo" [45], e são utilizadas em grafos de coautoria binários e não orientados.

A centralidade baseada no grau de um nó é definida como o número total de arcos que são adjacentes a este nó. A centralidade de grau representa a noção mais simples de centralidade, pois esta mede apenas quantas conexões ligam o autor aos seus vizinhos imediatos na rede.

No entanto, autores podem ser muito conectados com seus vizinhos imediatos mas ser parte de grupo isolado, um clique. Embora, estes autores possam ser bem conectados aos seus vizinhos imediatos, a sua centralidade de modo geral é baixa. Por outro lado, a centralidade de fluxo *closeness centrality* expande a definição de centralidade de grau focando em quão próximo um autor está de todos os outros. Para calcular a centralidade de fluxo de um vértice no grafo é preciso determinar os caminhos mínimos entre todos os vértices (autores) do grafo, e inverter estes valores para chegar ao valor da centralidade. Um autor central é caracterizado por ter muitos caminhos curtos aos outros autores na rede de coautoria.

Centralidade de autovetores ou PageRank

PageRank é o mecanismo que está no cerne do sistema de buscas da *Google*[6]. No *PageRank*, um *hyperlink* é considerado um relacionamento que atribui importância, ou avaliza ao que ele referencia. No PageRank a definição de prestígio é bem diferente da definição das medidas grau, proximidade, e centralidade de fluxo através do modelo de transferência ou herança de status.

Uma página possui alta posição na classificação se a soma das posições de seus *backlinks* (recomendações recebidas) é alta. O *PageRank* pode ser calculado usando um algoritmo iterativo simples, e corresponde ao principal autovetor da matriz de incidência das páginas da web normalizada.

O *PageRank* foi originalmente desenhado para encontrar os melhores resultados de busca na web baseado na estrutura dos seus *hiperlinks*, que é um grafo binário orientado por natureza. Assim, também é possível aplicar o algoritmo do *PageRank* ao modelo de redes de coautoria orientado de peso unitário.

2.4 Web of Science

O *Web of Science* foi lançado em 1997 e é uma interface voltada para navegadores de internet que provê acesso Web ao banco de citações ISI. Ele possui algumas funções a destacar: primeiramente, ele permite busca simultânea nos três bancos de citações do ISI. Isto permite a identificação de todos os autores citados na base de dados. O ISI possui um índice relacionando as referências aos artigos citados [47] [39].

2.4.1 Categorias de áreas temáticas de pesquisa

Nos experimentos foram utilizadas as **categorias de áreas temáticas de pesquisa (*Broad-Subject-Fields*)** obtidas no procedimento do *ranking* de universidades realizado pela ARWU (*Academic Ranking of World Universities*)[12], O ARWU usa a tabela de

categorias de pesquisa de periódicos disponível na página da *internet* do ISI [40].
A *Thomson Scientific* compilou as categorias de pesquisa de periódicos colecionados em dois índices:

SCIE Science Citation Index-Expanded

SSCI Social Science Citation Index

As tabelas de *Broad-Subject-Fields* indicadas abaixo classificam os artigos de periódicos em:

SCI Natural Sciences and Mathematics;

ENG Engineering/Technology and Computer Sciences;

LIFE Life and Agriculture Sciences;

MED Clinical Medicine and Pharmacy;

SOC Social Sciences;

INTER Interdisciplinary and Multidisciplinary Sciences.

2.5 Unificação de Objetos

O problema da ambiguidade na identificação de nomes de autores, instituições, publicações é um grande desafio para o trabalho de análise de dados bibliográficos [32]. Inicialmente foram estudados alguns trabalhos utilizando redes sociais para auxiliar nesta tarefa [32] [24]. Para que sejam realizadas análises de redes sociais em bases de dados é necessário um pré-processamento se os conjuntos de dados têm problemas de unicidade. Este é um dos problemas atacados por técnicas de limpeza de dados (dados faltantes, errôneos, duplicados e outros).

A solução desse tipo de problema pode ser formalizada como um processo de limpeza dos dados, geralmente aplicado à integração de várias bases de dados de fontes diferentes referentes às mesmas entidades do mundo real, um problema conhecido como *reference reconciliation* (*record linkage*, *merge/purge*, *de-duplication*, *hardening soft databases*, *reference matching*, *object identification* e *identity incertainty*) [9] [5] [17] [28].

O processo de unificação de objetos (*object consolidation*) [9] é um dos desafios da limpeza de base de dados, ou seja, os objetos numa base de dados são frequentemente representados por um conjunto de atributos, que sozinhos nem sempre identificam unicamente o objeto. O objetivo da unificação de objetos é corretamente (agrupar/ determinar)

todas as representações do mesmo objeto, para cada objeto num conjunto de dados. A unificação de objetos lida com desafios um pouco maiores que a maioria das técnicas de limpeza de dados, pois trata conjuntos de objetos com representações diferentes, onde grande parte dos atributos não está presente nas variedades de representações.

O problema da remoção de duplicações (*record deduplication* ou *record linkage*) tem um objetivo semelhante, mas lida com informações contidas em tabelas e conta apenas com os atributos do registro. No caso da unificação de objetos, no lugar de registros em tabelas, se lida com entidades, ou objetos, que são conceitos com uma semântica de nível mais alto de abstração. No caso da remoção de duplicações se assume que muitos atributos estão disponíveis em cada registro, o que torna o processo de remoção de duplicações efetivo. No entanto, na unificação de objetos, poucos atributos podem estar disponíveis em cada objeto o que torna o processo mais desafiador.

Existe uma variante particular da unificação de objetos na situação em que é conhecida uma lista dos possíveis objetos que podem ser encontrados no conjunto de dados, com os dados limpos, sem duplicações. Neste caso, o processo é chamado de reconciliação de suas referências conforme descrito em [17].

Relação com o desafio da ferramenta

Neste trabalho, as informações estão presentes numa rede de relacionamentos indiretos (coautoria, afiliação, o domínio do correio eletrônico do autor e o domínio dos endereços de correio eletrônico daquela universidade, ou do país, etc.).

Adicionalmente, estes problemas de ambiguidade normalmente surgem se um conjunto de dados foi construído a partir de fontes de diversas origens sem a preocupação da unicidade dos objetos, como é o caso deste trabalho que relaciona informações de duas fontes diferentes: dados bibliográficos da WOS e uma lista de nomes do quadro de professores de universidades de interesse.

Portanto, dois dos problemas fundamentais deste trabalho que devem ser formalizados são a unificação de um objeto e a reconciliação de suas referências.

Todos estes mecanismos são problemas de agrupamento por similaridade. Assim, foram estudadas abordagens para lidar com o problema de casamento de padrões e com agrupamento de objetos em conjunto de dados de grande volume.

Na base dados deste projeto pressupõe-se que os artigos são unicamente identificados, ou seja, eles possuem uma chave numérica única que os identifica (vide detalhamento destes dados nos anexos A.1 e B.1). Dentre as informações que são de interesse, é importante destacar que um artigo relaciona univocamente o título, palavras chaves, o nome abreviado do autor (às vezes existe associado a este nome abreviado o nome completo), os atributos do veículo, palavras chaves e área de pesquisa, e a contagem de citações. A base de dados utilizada é integralmente na língua inglesa.

2.5.1 Unificação de objetos utilizando relacionamentos entre os objetos

Nesta seção é apresentada uma abordagem para unificação de objetos baseada em similaridade de características e em informações extraídas de sua rede social na forma de análise de seus relacionamentos dispostos no seu *Attributed Relational Graph* (ARG) descrita no artigo [9].

A abordagem do artigo [9] mostrou-se bastante apropriada para o problema investigado, ou seja, o uso da estrutura de relacionamentos contida na rede social para agregar mais confiança nos resultados das análises da própria rede social. As técnicas tradicionais de limpeza de dados utilizam métodos baseados em similaridade das características dos objetos, alguns utilizam informações obtidas relações diretas com objetos chamadas de contexto e nesta abordagem além destas utiliza cadeia de relações entre os objetos conforme pode ser vista na Figura 2.2.

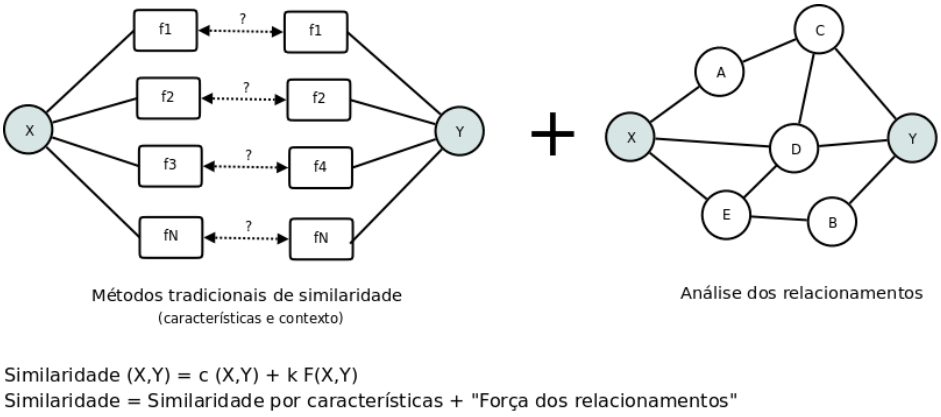

Figura 2.2: Similaridade de atributos considerando a "força dos relacionamentos" (baseado em [8]).

A abordagem é composta pelos passos abaixo:

- Construção do *Attributed Relational Graph* (ARG)

  O conjunto de dados é representado em um grafo não orientado chamado de *Attributed Relational Graph* $G = (V, E)$, onde os vértices V representam as entidades, ou objetos, do conjunto de dados e suas arestas E correspondem às relações entre as entidades. Os vértices podem representar objetos unificados ou aqueles que ainda não foram limpos. Entre os objetos não unificados de mesma classe existem relações de similaridade. O outro tipo de aresta representa as relações chamadas de regulares, que correspondem a relacionamentos entre as entidades (autoria entre

autores e artigo, afiliação entre autor e instituição, entre área de pesquisa e artigos). A construção do *ARG* é realizada através dos passos descritos abaixo:

  - Usando agrupamento por similaridade de atributos são obtidas arestas de similaridade entre os representantes de entidades de mesma classe (as similaridades entre entidades advêm da resultante de uma ou mais **relações de similaridade**). Apenas as similaridades de um valor mínimo t_0 ou maior são representadas.
  - As entidades são relacionadas entre si por **relações regulares** de associação entre autor e artigo, autor e universidade, artigo e universidades, etc.

- Calculo da força dos relacionamentos dentro de cada *Virtual Connected Subgraph* (*VCS*)

  Tipicamente o *VCS* é um componente conexo de um conjunto de *entidades da mesma classe* e suas *relações de similaridade* e as *associações de relações regulares* entre estas entidades, conforme pode ser visto na Figura 2.3. O cálculo da força dos relacionamentos é descrita nos passos a seguir:

  - Em cada *Virtual Connected Subgraph* (*VCS*) é calculada a força da conectividade entre (u, v) para cada u e v no *VCS*.
  - Para cada par de entidades (u, v) com relação de similaridade entre si é calculada a força da conectividade entre elas, *c(u,v)*. Desta forma, fortalecendo as relações de similaridade devido à força de conectividade deste grafo, ou seja, usando para isto relações entre u, v por outros caminhos diferentes da similaridade direta entre eles.

- A segmentação de cada *Virtual Connected Subgraph* com base nas relações de similaridade considerando a força de conectividade dos relacionamentos entre os objetos.

Apesar da importância desta abordagem para o problema de identificação de objetos baseada em similaridade de características e em informações extraídas de sua rede social, existem outros artigos relacionados a este problema dos quais ideias importantes foram utilizadas no trabalho discutidas no decorrer deste capítulo. Exemplos concretos desta abordagem podem ser vistos no anexo H.

2.6 Funções de similaridade

Métodos de casamento de padrões são utilizados para encontrar ocorrências de palavras chaves em um texto. Neste projeto, a busca pode não encontrar o padrão exato. Isto

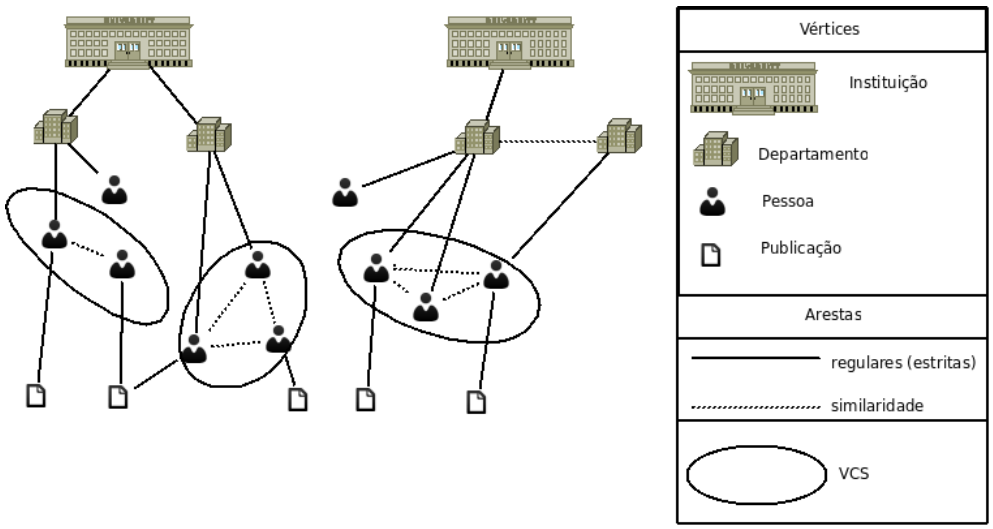

Figura 2.3: Virtual Connected Subgraph (VCS) (baseado em [8]).

ocorre porque os textos a serem trabalhados podem ter diversidade na grafia ou mesmo erros de caracteres. Normalmente em artigos digitalizados cujo texto foi reconhecido por software "OCR" este erro é comum. Mas, tendo em vista essa preocupação, o processo para encontrar as ocorrências de padrões nos textos deverá ser aproximado para admitir os possíveis erros e possibilitar a recuperação do que se pretendia expressar. A grafia de cada palavra chave (nomes de autores, afiliação, título, etc.) pode ser escrita de várias formas diferentes e podem ocorrer erros nos caracteres, então a todo padrão encontrado deverá ser associada uma forma canônica. A razão é que, se for considerada exatamente a informação que a busca por padrões obteve, isto implicaria em que a informação original representada por este padrão seria perdida. Assim, é importante manter a forma canônica e a original da informação. Além disso, a forma canônica também auxiliará na contagem de informações (quantas ocorrências casam com uma dada forma canônica, ou seja, com o objeto real que se busca representar). Neste texto, a forma canônica é denominada também como formato de referência, protótipo ou pivô [42] [4].

Mais detalhes sobre como estes conceitos e trabalhos se relacionam ao projeto encontram-se na seção 2.6.1. As nossas contribuições e a forma como estas ideias foram adaptadas ao projeto estão na seção 3.6.2 e as discussões relacionadas a esta questão estão nas seções 4.2.

2.6.1 A similaridade de nomes de autores

O artigo [36] reúne várias medidas de similaridade para textos em geral e para nomes. Mas, necessita-se identificar de forma aproximada a similaridade entre dois nomes. O ideal é que

a função de similaridade seja capaz destacar o que é completamente diferente do que é um pouco parecido, e que seja barata o suficiente para que todos os nomes sejam comparados entre si. Neste caso, apenas é feito um processamento prévio identificando os nomes exatamente iguais, e então, se parte deste conjunto para fazer os cálculos de similaridade, mas é mantida a informação do conjunto de representações que estão representadas pelo mesmo nome. Este pré-processamento é uma ordenação de elementos através do algoritmo *insert-sort* usando mecanismos de comparação via sua representação *hash*, sendo assim, de complexidade $O(n \log(n))$, onde n é o número de elementos a serem ordenados.

O cálculo da similaridade é baseado na distância entre textos chamada de *bag of words* [36], no qual se prepara um vetor binário onde cada índice corresponde a uma palavra do vocabulário de palavras existentes em todos os textos do conjunto de dados, indicando a existência ou não desta palavra naquele texto.

A distância entre dois textos é o número de palavras em comum, ou seja, o produto interno dos seus respectivos vetores (*bag of words*).

Este cálculo pode ser aplicado analogamente para similaridade entre denominações de entidades como: nomes de autores, nomes abreviados de autores, nomes de veículos de publicação, nomes de cidades, nomes de estados. Este tipo de tarefa envolve elementos que possuem poucas palavras, algo da ordem de 2 ou 3, no pior caso uma dezena de palavras. Sendo assim, foi proposto uma variação deste método considerando os bigramas destas frases (seção 3.4).

2.6.2 Similaridade de textos usando Função *Kernel*

Esta categoria de cálculo de similaridade entre textos usa uma função *kernel* para calcular a proximidade entre dois textos [30]. A ideia é comparar o texto padrão com o texto de entrada pelas suas subsequências. Os textos terão uma maior similaridade quanto maior o número de subsequências iguais. Os caracteres que compõem uma subsequência não precisam estar todos contíguos no texto alvo para que seja acusado que ela foi encontrada, e nesse caso usa-se um peso para determinar o grau de contiguidade. A subsequência "sem", por exemplo, está presente na palavra "**sem**pre" e na palavra "s**em**", mas com pesos diferentes.

Neste algoritmo, cada texto é mapeado em um conjunto de características formando um "espaço de características". Cada subsequência existente fornece uma dimensão neste espaço e o valor desta coordenada depende de quão frequente e compacto é aquela subsequência no texto. Para subsequências não contíguas utiliza-se um fator de decaimento $\lambda \in (0, 1)$ que é usado para "pesar" ou medir a presença de uma característica no texto. Para comparar dois textos usa-se uma função *kernel* que é uma função que calcula o produto interno entre os dois espaços de características.

A função kernel definida por Huma Lodhi *et* al. [30] é dada por:

$K'_0(s,t) = 1$, para qualquer s e t;
$K'_i(s,t) = 0$, se $min(|s|,|t|) < i$;
$K_i(s,t) = 0$, se $min(|s|,|t|) < i$;
$K'_i(sx,t) = \lambda K'_i(s,t) + \sum_{j:t_j=x} K'_{i-1}(s,t[1:j-1])\lambda^{|t|-j+2}$;
$K_n(sx,t) = K_n(s,t) + \sum_{j:t_j=x} K'_{n-1}(s,t[1:j-1])\lambda^2$;

A figura 2.4 ilustra um exemplo do uso da função *kernel*. Nela são considerados as palavras "cat", "car" e um tamanho de subsequência igual a 2. Isto gera um espaço de características de tamanho 5. Os valores de cada característica foram calculados de acordo com a função definida acima. Dessa forma para encontrar a semelhança entre as palavras "cat" e "car" realiza-se um produto interno dos dois vetores de características, portanto $K_2(cat,car) = \lambda^4$, pois somente a coordenada referente à subsequência "ca" possui um valor diferente de zero referente a ambas as palavras. Este resultado não está normalizado, para isso deve-se obter $K_2(cat,cat) = K_2(car,car) = 2\lambda^4 + \lambda^6$ e normalizar dividindo λ^4 por $2\lambda^4 + \lambda^6$. De modo geral, a semelhança normalizada de dois textos s e t é dada por:

$$K^*_2(s,t) = \frac{K_2(s,t)}{\sqrt{K_2(s,s)K_2(t,t)}};$$

Na figura 2.4 são apresentadas as palavras "cat" e "car" (duas subsequências de tamanho 2) que possuem vetores de características definidos por $\phi(cat)$ e $\phi(car)$. O cálculo do valor das coordenadas dos vetores de características $\phi(cat)$ e $\phi(car)$ utiliza a função *kernel* definida.

Os experimentos realizados por Huma Lodhi *et al.* mostram que este algoritmo pode ser uma alternativa eficiente para casamento aproximado de textos. O problema encontra-se no tempo necessário para computar a função *kernel* e na definição do tamanho da subsequência e o valor de λ, que são parâmetros a serem definidos de acordo com a aplicação e características dos textos comparados.

	ca	ct	at	cr	ar
$\phi(cat)$	λ^2	λ^3	λ^2	0	0
$\phi(cat)$	λ^2	0	0	λ^3	λ^2

Figura 2.4: Exemplo de similaridade por string kernel.

Esta função foi implementada dentro das atividades do projeto *Ranking de Publicações baseado na Extração de Textos da Internet* [15] e foi utilizado como parte da função *compare authors*. A função *compare authors* é utilizado pelo algoritmo *bipartite*, na pré-seleção dos artigos na *fase 1* da metodologia 4.3 e na fase final do agrupamento 4.2.

2.7 Algoritmos de agrupamento (Clustering)

Sistemas de classificação de objetos são utilizados para categorizar objetos segundo suas características. Estes mecanismos podem utilizar uma abordagem vise à utilidade para agrupar os objetos ou uma abordagem exploratória que permite que os agrupamentos emerjam naturalmente. Estas duas abordagens principais são definidas abaixo:

classificação supervisionada neste tipo de classificação existe um mapeamento direto de um vetor que descreve as características de cada objeto do conjunto a ser a categorizado e um vetor finito de categorias. Um algoritmo de aprendizado é utilizado identificar os parâmetros deste mapeamento de forma otimizada baseado em exemplos, buscando minimizar os riscos deste processo empírico.

classificação não supervisionada este tipo de categorização também é chamada de agrupamento, ou análise exploratória. Neste tipo de abordagem não há dados previamente categorizados. O objetivo do agrupamento é separar um conjunto finito de objetos não categorizados em conjuntos finitos e discretos "naturais", com base em estruturas que não são explicitas.

Algoritmos de agrupamento dividem os dados em certo número de agrupamentos, de forma que haja homogeneidade interna nos agrupamentos e separação externa, com base em algum tipo de definição de similaridade entre objetos, objetos e grupos e entre grupos.

O conjunto de todos os agrupamentos encontrados pode ser de vários tipos em relação a algumas propriedades: aninhamento, sobreposição e completude. A completude considera se o agrupamento classificou todos os objetos ou parcialmente. A sobreposição avalia se os objetos pertencem exclusivamente a um agrupamento ou podem pertencer a mais de um ao mesmo tempo. Quando se permite que o objeto esteja em mais de um agrupamento isto pode ocorrer de forma absoluta ou nebulosa (*Fuzzy*).

O resultado do agrupamento pode obtido na forma de classes aninhadas, onde as classes mais internas são formadas por elementos mais similares, formando uma estrutura hierárquica. Assim, quando se obtém agrupamentos aninhados processo agrupamento é chamado de hierárquico e quando são obtidas divisões rasas de particional, raso ou não hierárquico. A essa estrutura hierárquica de classes resultado de um agrupamento hierárquico se dá nome de dendograma.

Para formalizar os algoritmos de agrupamento podemos definir o conjunto de representações de objetos a ser classificado R. Onde $R = \{r_1, \ldots, r_j, \ldots, r_N\}$ e $r_j = (r_{j1}, r_{j2}, \ldots, r_{jd})_T \in \mathcal{A}^d$. Sendo que, cada medida r_{ji} corresponde ao atributo i da representação de um objeto j. Ou seja, cada objeto é representado por um vetor de características de dimensão d em $\mathcal{A}^d$.

- Agrupamento não hierárquico busca encontrar uma divisão em K partes do conjunto R, $C = \{C_1, \ldots, C_K\}(K \leq N)$, de forma que:

  1. $C_i \neq \emptyset$, $i = 1, \ldots, K$;
  2. $\cup_K^{i=1} C_i = R$;
  3. $C_i \cap C_j = \emptyset$, $i, j = 1, \ldots, K$ e $i \neq j$.

- Agrupamento hierárquico busca construir uma estrutura aninhada na forma de árvore (dendograma) com os elementos do conjunto R, $H = \{H_1, \ldots, H_P\}(P \leq N)$, de forma que se $C_i \in H_m$, $C_j \in H_l$, e $m > l$ implica que $C_i \subset C_j$ ou $C_i \cap C_j = \emptyset$ para todos i, j. Com $j \neq i$ e $m, l = 1, \ldots, P$.

Geralmente o agrupamento hierárquico é utilizado quando é necessário obter uma relação hierárquica nos dados. No entanto, pode-se observar que a cada nível l do dendograma ($l = 1, \ldots, P$) há uma divisão de agrupamento na forma particional, e quando há uma definição heurística do que seria a melhor divisão em grupos, cada nível é analisado para se decidir qual é o nível que melhor representa esta definição de agrupamento.

Um agrupamento hierárquico pode ser obtido através de duas abordagens:

Aglomerativa esta abordagem de forma simplificada pode ser descrita como um processo que vai das partes para o todo. Ou seja, partindo de cada instância dos objetos sendo classificados como sendo um grupo. E a cada passo é feita a união dos grupos mais próximos. Para isso, é necessário definir a distância entre dois grupos. Sendo assim, no decorrer do texto serão apresentadas algumas definições de distância entre grupos.

Divisória esta abordagem poderia ser descrita como um processo que parte do todo e chega até as partes. Então apenas um grupo com todas as instâncias de objetos sendo classificados é o início do processo. A cada passo um agrupamento é dividido até que só restem grupos com apenas um objeto. Neste caso é necessário definir como será a escolha de qual grupo deve ser dividido a cada passo e como será feita esta divisão.

Distância entre dois grupos

Há uma variedade de formas diferentes de definir quais são os dois grupos mais próximos:

Single-link A similaridade máxima entre pares de grupos é definida pelo par de grupos que tiver a menor distância entre os elementos. O par de elementos com a menor

distância é composto por um elemento de cada grupo do par de grupos. Com essa definição de similaridade o processo de agrupamento pode resultar em grupos "longos" e "finos" devido ao efeito de encadeamento.

Complete-link A similaridade máxima entre pares de grupos é definida pelo par de grupos que tiver a menor distância entre os seus elementos mais distantes. O par de elementos com a maior distância é composto por um elemento de cada grupo do par de grupos. Com essa definição de similaridade o processo de agrupamento pode resultar em grupos "compactos" e "esféricos".

Centroid A similaridade máxima entre pares de grupos é definida pelo par de grupos que tiver a menor distância entre o ponto médio de seus elementos (centroides). Em cada grupo do par de grupos sendo comparado é calculado o ponto médio (centroide) entre seus elementos.

Averange-link A similaridade máxima entre pares de grupos é definida pelo par de grupos que tiver a menor distância média entre os seus elementos. Os pares de elementos utilizados para o cálculo da distância média é composto por um elemento de cada grupo do par de grupos.

Group-averange A similaridade entre dois grupos é dada pela média de todos os pares do grupo resultante da união destes dois grupos.

Ward Cada grupo é representado por seu centroide (média entre as distâncias entre todos os elementos do grupo). A medida de proximidade entre dois grupos é definida pelo menor aumento do desvio padrão quando dois grupos são unidos. O método de Ward busca minimizar a soma dos quadrados das distâncias entre os objetos e os centroides de seus grupos [44].

2.7.1 Algoritmos de agrupamento baseados em teoria dos grafos

Os problemas de agrupamento podem, de forma bastante conveniente, utilizar conceitos e propriedades que tem paralelo na teoria dos grafos. Podemos descrever proximidades entre objetos em termos de pesos nas arestas E em um grafo G onde os vértices V correspondem aos objetos a serem classificados. O grafo pode ser simplificado considerando apenas as arestas que possuam uma dissimilaridade mínima t_0 com base numa matriz de dissimilaridade entre os objetos pode ser definida como $D_{i,j} = 1$ somente se $D(x_i, x_j) \leq t_0$ e zero caso contrário.

Neste grafo simplificado podem ser utilizadas várias definições de agrupamento como subgrafo maximal conexo *single link*, ou subgrafo completamente conexo *clique*, e variações entre estes estremos como o método Ward [48].

O processo pode ser divisivo ou aglomerativo e podem ser utilizadas heurísticas para direcionar o processo de unir agrupamentos ou para dividi-los.

2.7.2 Algoritmo de agrupamento baseado em similaridade estrutural

O artigo de Pascal Pons [38] tem o objetivo de encontrar agrupamentos de objetos representados num grafo não orientado $G = (V, E)$ utilizando medidas de similaridade estrutural neste grafo e descreve o problema em termos de um algoritmo de agrupamento aglomerativo baseado no método Ward.

Este grafo descreve a similaridade entre dois objetos pelo número de arestas que há diretamente entre eles, por definição todo objeto é similar a si mesmo. Assim, a similaridade é definida em termos da matriz de transição P de um processo de passeio aleatório de comprimento t neste grafo. Onde a probabilidade P_{ij} de passar do vértice i para o j depende do número de arestas que existem entre os vértices i e j e o número de arestas que partem deste vértice i (grau $g(i)$ do vértice de partida):

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(P_{ik}^t - P_{jk}^t)^2}{g(k)}} \tag{2.1}$$

A distância entre duas comunidades (grupos) C_1 e C_2 ($r_{C_1C_2}$) leva em conta passeios aleatórios que iniciam em uma comunidade (o vértice de partida dentro da comunidade é escolhido aleatoriamente e uniformemente entre todos os vértices desta comunidade). Desta forma é definida a probabilidade de sair da comunidade C em direção ao vértice j em t iterações como:

$$P_{C_j}^t = \frac{1}{|C|} \sum_{i \in C} P_{ij}^t$$

Considerando $C_1, C_2 \subset V$ sendo duas comunidades, a distância $r_{C_1C_2}$ entre estas duas comunidades é definida por:

$$r_{C_1C_2} = \sqrt{\sum_{k=1}^{n} \frac{(P_{C_1k}^t - P_{C_2k}^t)^2}{g(k)}} \tag{2.2}$$

O algoritmo inicia com uma divisão de agrupamentos P_1 do grafo em n grupos (ou

comunidades) reduzidos a um único vértice. As distâncias são calculadas entre todos os vértices adjacentes. Então o mecanismo de agrupamento evolui seguindo os seguintes passos. A cada iteração k:

- duas comunidades C_1 e C_2 são escolhidas em P_k de acordo com um critério baseada na distância entre as comunidades (método Ward).

- as duas comunidades são unidas em uma nova comunidade $C_3 = C_1 \cup C_2$ e uma nova divisão é criada: $P_{k+1} = (P_k \setminus C_1, C_2) \cup C_3$, e

- as distâncias entre as comunidades são atualizadas

  Depois de $n-1$ iterações o algoritmo termina e obtém $P_n = V$. Cada iteração define uma divisão P_k do grafo em comunidades, que resulta numa estrutura hierárquica chamada de dendograma.

O algoritmo induz uma sequência $(P_k)_{1 \leq k \leq n}$ de partições em comunidades. Para saber qual destas partições captura a estrutura de similaridade denominada de comunidade é utilizado o critério da modularidade Q. A modularidade Q considera a quantidade de arestas dentro da comunidade C (e_C) em relação as que saem da comunidade C (a_C):

$$Q(P) \quad = \quad \sum_{C \in P} e_C - a_C^2$$

A melhor partição é a que maximiza Q.

Essa abordagem tem a eficiência $O(n^2 \log n)$ em tempo e $O(n^2)$ em espaço para grafos esparsos ($e = O(v)$), para encontrar conjuntos de nós densos em grafos esparsos.

Os agrupamentos nesta abordagem são compostos de vértices (representantes), que estão no mesmo componente conexo do grafo (vide item 2.7 sobre o conceito de algoritmo de agrupamento e a abordagem de unificação de objetos em 2.5). Nestes algoritmos um desafio potencial é o de lidar com grafos de milhares de vértices. Por isso é importante segmentar o grafo original em componentes conexos.

Caso o grafo não seja esparso há a alternativa do algoritmo apontado em [10] que é $O(md \log n)$ em tempo, sendo d a profundidade do "dendrograma" que representa a estrutura da comunidade. No entanto ele tem o vício de gerar agrupamentos de tamanhos semelhantes [38].

2.7.3 Algoritmos de agrupamento e o problema de unificação de objetos

O problema de unificação das representações dos objetos pode ser formalizado como um problema de agrupamento (*Clustering*) que visa identificar grupos cujos elementos tenham o máximo de similaridade entre si e o mínimo de similaridade com elementos fora do grupo.

Ou seja:

$O = \{O_i, \ldots, O_{|O|}\}$ é o conjunto de entidades do mundo real, e $|O|$ é desconhecido;

$X = \{X_i, ..., X_{|X|}\}$ é o conjunto de representações destes objetos;

$d[X_i]$ corresponde à entidade O_i ao qual X_i se refere e esta relação é desconhecida inicialmente;

$C[O_i]$ é o agrupamento de todas as representações que se referem à mesma entidade de $d[X_i]$. O "agrupamento" é algo desconhecido inicialmente, o objetivo do algoritmo de agrupamento é encontrar o grupo ao qual pertence cada X_i;

$S[O_i]$ corresponde a todas as representações que poderiam ser associadas a O_i, onde $C[O_i] \subset S[O_i]$. $S[O_i]$ corresponde ao "conjunto de unificação" que é geralmente determinado por similaridade entre seus atributos correspondentes dos objetos.

Algoritmos de agrupamento estritos particionam os dados X_i em certo número de agrupamentos (grupos, subconjuntos ou categorias) $S[O_i]$. Neste trabalho não são abordados agrupamentos difusos onde um objeto poderia pertencer a mais de um grupo ao mesmo tempo.

No modelo de algoritmo de agrupamento adotado as similaridades significativas entre as características de mesma classe dos representantes dos objetos são arestas de similaridade que ligam estas características. Os agrupamentos são compostos de vértices (representantes) que estão no mesmo componente conexo do grafo composto por todos os objetos de mesma classe que o algoritmo deve particionar. Esta abordagem é chamada de agrupamento baseado em teoria dos grafos [42] [48].

2.7.4 Comparação entre algoritmos de agrupamento

O algoritmo de agrupamento deste projeto utiliza um mecanismo para identificação de comunidades [38] como critério para definir um agrupamento (item 2.7.2) e este algoritmo tem um parâmetro de ajuste que limita o tamanho do passeio aleatório (*walk size*). O autor deste mecanismo sugere valores entre 4 e 5 para grafos esparsos, que é o caso dos

dados deste projeto. Mas era necessário comparar execuções do algoritmo de agrupamento como um todo vendo o impacto de variar este valor. Para isso era importante ter uma forma de comparar dois resultados de agrupamento.

A tarefa de comparar os agrupamentos resultantes de variações de algoritmos de agrupamento foi baseada no artigo [50]. Desta forma foi implementada uma ferramenta que usa o conceito de "*Graph-based cohesion*" e "*Graph-based separation*", e também levam em consideração algumas discussões de *Cluster Analysis* do livro *Introduction to Data Mining* [42].

Um conceito importante para entendimento deste mecanismo é o de componente conexo em um grafo de similaridade. Um componente conexo é composto de todos os elementos que são similares entre si ou de elementos similares a outro similar a estes.

***Graph-based cohesion*:** somatório da proximidade entre os elementos do mesmo agrupamento.

***Graph-based separation*:** somatório da proximidade entre os elementos que estão em agrupamentos diferentes (mas estão no mesmo componente conexo).

Cada algoritmo difere em como particiona estes componentes conexos do grafo de similaridade. Então *Graph based cohesion* e *Graph-based separation* são calculados para as partições dos componentes conexos propostos por cada algoritmo, e também do componente como um todo (que é um limitante superior da coesão).

Assim, compara-se a diferença no particionamento de cada componente conexo valorizando a *Graph-based cohesion* e depreciando a *Graph-based separation*. Quanto maior for o valor resultante melhor a coesão. Considerando o somatório destes valores de cada componente conexo é possível avaliar a qualidade dos agrupamentos como um todo para cada algoritmo.

Se forem utilizados os mesmos dados e as mesmas funções de similaridade, os componentes conexos obtidos para um mesmo limite de similaridade aceitável são iguais independentemente do algoritmo de agrupamento utilizado.

A ferramenta de comparação de algoritmos de agrupamento desenvolvida neste trabalho apresenta os resultados em termos do valor total do critério de qualidade e apresenta as diferenças (em termos do que foi adicionado ou removido e o que ficou igual) entre os agrupamentos obtidos por duas versões de algoritmo de agrupamento para cada componente conexo do grafo de similaridade.

2.7.5 Medidas padrão de desempenho em algoritmos de agrupamento

A medida do desempenho de mecanismos de classificação na área de Recuperação de Informação baseada em texto com múltiplas categorias geralmente é realizada com base em medidas de desempenho para cada categoria ou classe. A resposta do algoritmo é avaliada em relação a uma amostra de dados que foi previamente categorizada.

A atividade de atribuição de categorias a textos usando categorização binária pode ser avaliada usando uma tabela de contingência 2x2 para cada categoria[49]. Vamos considerar, sem perda de generalidade, que cada categoria corresponda a uma denominação de pesquisador constante em uma lista de referência (lista de autores pivôs). Além disso, que algumas denominações de autores de artigos são categorizadas segundo esta lista de referência. Estas informações são relacionadas na Tabela de contingência (Tabela 2.3) para um elemento da lista de referência (autor pivô).

Tabela 2.3: Tabela de contingência para a identificação de um autor pivô

	o correto é SIM	o correto é NÃO
Atribuiu SIM	a	b
Atribuiu NÃO	c	d

A tabela de contingência para identificação de um autor pivô possui quatro células (a,b,c,d) descritas abaixo:

a: uma denominação foi atribuída corretamente a um elemento de referência;
b: uma denominação foi atribuída incorretamente a um elemento de referência;
c: a denominação foi incorretamente rejeitada como sendo deste elemento de referência;
d: a denominação foi corretamente rejeitada como sendo deste elemento de referência.

As medidas de desempenho padrão para Recuperação de Informação em Textos podem ser calculadas dessas tabelas de contingências. Essas medidas são recuperação ou *recall*(R), precisão ou *precision*(P), falsos positivos ou *fallouts*(f), acurácia ou *acuracy*(A) e erro (E). As respectivas fórmulas são listadas abaixo:

$R = a/(a+c)$ se $a+c>0$, caso contrário é indefinido;
$P = a/(a+b)$ se $a+b>0$, caso contrário é indefinido;
$f = b/(b+d)$ se $b+d>0$, caso contrário é indefinido;
$A = (a+d)/n$ onde $n = a+b+c+d>0$;
$E = (b+c)/n$ onde $n = a+b+c+d>0$.

Para avaliar o desempenho médio de todas as categorias, há dois métodos convencionais denominados de média-macro-agregada (*macro-averanging*) e média-micro-agregada (*micro-averanging*). O método de médias-macro-agregadas das medidas de desempenho é calculado computando as medidas acima para cada tabela de contingência de cada categoria e depois computando a média das medidas de todas as categorias. O método de médias-micro-agregadas das medidas de desempenho é calculado a partir de uma tabela de contingência consolidada de todas as tabelas de contingência de cada categoria, ou seja, cada célula (a,b,c,d) da tabela de contingência de cada categoria irá contribuir para o valor da tabela de contingência consolidada ($\sum a$,$\sum b$,$\sum c$,$\sum d$). O valor da média-micro-agregada das medidas de desempenho é calculado a partir desta tabela de contingência consolidada e dá igual peso para qualquer denominação que está sendo categorizada. É considerada, então, uma média por denominação sendo categorizada (ou seja, uma média em relação a todos os pares denominações/professores). Por outro lado, o valor da média-macro-agregada das medidas de desempenho concede peso igual a cada categoria, independentemente de sua frequência, ou seja, uma média por categoria. No caso do experimento realizado neste trabalho será utilizada a média-micro-agregada, pois ela destaca o desempenho em categorizar denominações.

Quando são comparados dois algoritmos de categorização é desejável ter uma única medida de eficácia. A medida F_β definida por van Rijsbergen [49], é comumente utilizada com este intuito, balanceando a recuperação e a precisão por um parâmetro β. No caso mais comum essa medida é chamada de F_1 que concede igual importância para a recuperação e a precisão:

$$F_1 \;=\; 2PR/(R+P)$$

Na sua forma mais geral a medida F_β é definida como:

$$F_\beta \;=\; \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

O parâmetro β permite diferenciar o peso da precisão(P) e da recuperação(R). Por exemplo, $\beta = 0,5$ atribui o dobro de peso à precisão.

O mecanismo de unificação de objetos do presente trabalho (item 3.5.2) identifica dentre as denominações de autores nos artigos as denominações que correspondem a um mesmo autor. Alguns autores também são identificados com as denominações de uma lista de referência fornecida ao sistema (autores pivôs). Isto pode ser visto como um problema de categorização de textos com múltiplas categorias, onde cada autor pivô (elemento da

lista de referência) é uma categoria. No mecanismo de unificação de objetos as categorias são os membros do quadro de professores das instituições acadêmicas dados como entrada ao sistema.

2.8 *Rankings* baseados em desempenho em pesquisa

Desde há alguns anos *rankings* de universidades e publicações têm sido produzidos tomando por base os dados contidos em bases de dados bibliográficas como os da WOS ISI (seção 2.4) ou através de uma minuciosa pesquisa de opinião. Atualmente, dentre os mais conhecidos em *rankings* de universidades e instituições de pesquisa há o ARWU (*Academic Ranking of World Universities*)[12], Times[18], SCIMago[19] e USNews[31]. Já em *rankings* de publicações há o JCR (seção 2.4).

O ARWU utiliza dados de publicações produzidos pelos membros do quadro efetivo de 2000 instituições acadêmicas do mundo todo. Esses dados correspondem à produção do ano anterior, e são agrupados por *Broad-Subject-Fields*, conforme descrito na seção 2.4. Em 2009, passou a utilizar também outro critério de agrupamento por área de pesquisa. Essa definição de área de pesquisa que difere da definição proposta para a ferramenta, baseada em categorias de pesquisa (anexo C). No caso da ARWU é adotada a definição área de pesquisa da ISI (seção 2.4) usada para agrupar os pesquisadores mais citados. A definição de áreas de pesquisa da ISI é baseada no periódico onde foi publicado o artigo. Para cada área de pesquisa existe uma lista de periódicos que a representa, dividas em 21 categorias (*Computer Sciences*, *Mathematics*, etc.). Os outros critérios da ARWU para avaliar o prestígio em pesquisa científica são bastante discutíveis mas têm intenção de utilizar dados disponíveis publicamente para identificar a qualidade das instituições como premiações em cada campo de pesquisa, número de pesquisadores na lista dos mais citados na ISI, e número de artigos nas revistas *Science* e *Nature*.

A USNews para avaliar os melhores programas de PhD das universidades norte americanas utiliza pesquisas de opinião. As universidades participantes da pesquisa são as que tiveram pelo menos cinco conclusões de doutorado de 2003 a 2008 de acordo com o relatório da *National Science Foundation - Science and Engineering Doctorate Awards*. Os questionários são enviados aos chefes de departamento e aos coordenadores de cada programa em cada disciplina. Os entrevistados devem dar notas aos programas de cada instituição de "fraco" (1) a "excelente" (5). Somente as instituições que obtiverem nota média mínima de 2 é que participam da classificação. A pesquisa de opinião sobre as áreas científicas biologia, química, ciência da computação, ciências da terra, matemática, física, e estatística ocorreram durante o outono de 2009.

OS *rankings* ARWU e USNews são comparados com um experimento no qual o objetivo é capturar dentro da estrutura da rede social a informação de prestígio e atribui-la às

instituições acadêmicas aos quais são afiliados os professores participantes da rede (seção 4.5).

2.8.1 Comparando *rankings*

O coeficiente de correlação entre *rankings Kendall tau* avalia a proporção em que dois *rankings* são similares, levando em conta apenas os participantes em comum[41]. Considerando os *rankings* R e S que são comparados. E seja I o conjunto dos elementos i participantes destes dois *rankings*. Onde as posições dos elementos i nos *rankings* R e S são denotadas respectivamente por r_i e s_i. E assumindo sem perda de generalidade um dos *rankings* como referência, que será o *ranking* R.

O princípio em que se baseia o coeficiente de *Kendall* é que se há associação entre os *rankings* e estando os elementos participantes I na ordem das posições r_i no *ranking* R (I_R), os valores das posições s_i do *ranking* S apresentarão também uma tendência crescente se houver associação positiva entre os *rankings*. E no caso de haver tendência decrescente nos valores das posições s_i do ranking S indicará que há uma relação for inversa entre os *rankings*.

Kendall propôs que após considerar os participantes na ordem de posição de um dos *rankings*, pode-se contabilizar os elementos que estão em ordem concordante e discordantes nos dois *rankings*, em relação à máxima concordância possível. Abaixo é apresentada a contagem do número de pares concordantes menos a quantidade de pares discordantes, também chamado de *Score(S)*:

$$nc - nd \;\; = \sum_{i,j \in I_x \quad \text{e} \quad i<j} \text{sign}(r_j - r_i) \times \text{sign}(s_j - s_i)$$

Onde:

$$\text{sign}(x) = \begin{cases} +1 & \text{se } x > 0 \\ 0 & \text{se } x = 0 \\ -1 & \text{se } x < 0 \end{cases}$$

Na situação onde não há elementos com mesma posição nos *rankings*, ou seja, não há elementos participantes de um *ranking* que receberam a mesma classificação naquele *ranking* (participantes em "empate"), então a máxima concordância possível (D) é a quantidade de combinações de pares de participantes.

Se $n = |I|$, então:

$$D = \binom{n}{2}, \text{ o máximo valor possível para o } \textit{Score(S)}$$

O coeficiente τ de *Kendall* é a contagem de concordantes menos os discordantes em relação à máxima concordância possível:

$$\tau = S/D \tag{2.3}$$

Este coeficiente também pode ser visto como a proporção de concordantes menos a proporção de discordantes. Sendo assim $\tau \in [-1, +1]$, onde -1 é a discordância máxima, $+1$ corresponde à concordância máxima e 0 quando não há associação entre os *rankings*.

Para que a fórmula acima mantenha a propriedade do coeficiente $|\tau|$ de ter valor máximo 1 a quantidade de concordâncias máximas possíveis precisa descontar os "empates" de cada *ranking*. Desta forma o coeficiente de correlação passa ser denominado de τ_b. Se há n_x "empates" de extensão t_l, com $l = 1, \ldots, n_x$ dentro do *ranking* x e n_y "empates" distintos de extensão u_l, com $l = 1, \ldots, n_y$ dentro do *ranking* y, então:

$$D = \sqrt{(\binom{n}{2} - T)(\binom{n}{2} - U)}, \tag{2.4}$$

onde

$$T = \frac{1}{2}\sum_{l=1}^{n_x} t_l(t_l - 1),$$

e

$$U = \frac{1}{2}\sum_{l=1}^{n_y} u_l(u_l - 1).$$

Para realizar o teste de significância é necessário o cálculo da distribuição de probabilidade para cada valor possível de S, para todas as combinações de dois *rankings* de tamanho n. Isto geralmente é feito para $n \leq 10$. No entanto, para $n > 10$ e sem "empates" em ambos os *rankings* a distribuição de probabilidade de S converge para uma distribuição normal com média zero. Assim, o nível de confiança α é dado pela probabili-

dade de τ ser zero simplesmente devido ao acaso, ou seja a hipótese nula $\mathcal{H}_0$. Mas o teste da hipótese nula $\mathcal{H}_0$: $\tau = 0$ é equivalente a testar $\mathcal{H}_0$: $S = 0$.

Sendo assim quando $n > 10$ e não há "empates" nos *rankings* a variância de S é aproximada pela fórmula abaixo:

$$\sigma_S{}^2 = \frac{1}{18}n(n-1)(2n+5) \tag{2.5}$$

Quando $n > 10$ e há empates foi proposto por Valz, McLod e Thompson [43] uma derivação aproximada da fórmula da variância para o caso geral:

$$\begin{aligned}\sigma_S{}^2 = \{&\frac{1}{18}n(n-1)(2n+5) - \sum_{l=1} l = n_x t_l(t_l-1)(2t_l+5) - \sum_{l=1} l = n_y u_l(u_l-1)(2u_l+5)\} \\ &+\frac{1}{9n(n-1)(n-2)}\{\sum_{l=1} l = n_x t_l(t_l-1)(t_l-2) - \sum_{l=1} l = n_y u_l(u_l-1)(u_l-2)\} \\ &+\frac{1}{2n(n-1)}\{\sum_{l=1} l = n_x t_l(t_l-1) - \sum_{l=1} l = n_y u_l(u_l-1)\}\end{aligned} \tag{2.6}$$

Então para $n > 10$, hipótese nula pode ser obtida transformando S em um valor de Z:

$$Z_S = \frac{S}{\sqrt{\sigma_S{}^2}} \tag{2.7}$$

ou

$$Z_\tau = \frac{\tau}{\frac{\sigma_\tau}{\sqrt{D}}} \tag{2.8}$$

Ao final basta obter a probabilidade cumulativa da distribuição normal com média zero e variância um para o valor $-|Z_\tau|$ e multiplicar por 2. Se o valor for menor que o nível de confiança (α) desejado (tipicamente 5% ou 10%) é possível aceitar que os *rankings* estão correlacionados.

Capítulo 3

Ferramenta para análise de redes sociais em dados bibliográficos

Este capítulo descreve a ferramenta proposta para análise de redes sociais em dados bibliográficos. Essa ferramenta incorpora as contribuições do trabalho para resolver/mitigar o problema de ambiguidade na denominação de entidades (autores, afiliação, veículos e outros atributos que possam ter variações desconhecidas nos seus nomes), metodologias de agrupamento destas várias denominações que representam certa entidade e funções de similaridade entre nomes destas entidades. Essas contribuições são destacadas no decorrer da descrição da ferramenta. As discussões sobre os resultados obtidos e os desafios encontrados estão no capítulo 5.

3.1 Visão geral da abordagem utilizada

A abordagem adotada pela ferramenta para análise de redes sociais em dados bibliográficos é mitigar o problema de ambiguidade na identificação dos autores e instituições as quais os autores estão afiliados, e então realizar as análises dos dados. As instituições foram identificadas com expressões regulares nas consultas da extração dos dados bibliográficos da base de dados (veja mais detalhes na seção 3.2) e estamos desprezando o erro nesta etapa. O grande desafio se encontra na identificação dos autores. Esse problema é tratado utilizando a unificação de objetos descrita na seção 2.5.

Dada um área de pesquisa específica, a ferramenta objetiva realizar a análise de redes sociais em dados bibliográficos relacionados a esta área, em relação a um grupo de pesquisadores que compõe o quadro de professores dos departamentos/institutos desta área, em instituições de maior renome. Desta forma, a unificação de objetos pode considerar uma lista de objetos referência e assim adotar a sua variante chamada reconciliação de referências. No entanto, para que a rede social de coautoria não seja comprometida, é

importante considerar os coautores que não fazem parte da lista de autores de referência. Ainda que grande parte dos autores que fazem parte da lista de referência sejam os que têm maior relevância na rede de coautoria, os outros autores são agrupados com base em informações dos próprios dados.

Definido o problema dessa forma fica claro que esse é um problema característico de algoritmos de agrupamento, onde os grupos devem emergir com base nas informações dos dados, e também não se sabe quantos grupos devem ser encontrados. A lista de objetos de referência será agrupada juntamente com os autores e irá influenciar a divisão dos agrupamentos de forma que um agrupamento só contenha objetos de referência muito semelhantes entre si.

A menos desta pequena variação foram adotados os mesmos passos da unificação de objetos utilizando relacionamentos entre os objetos (seção 2.5.1). O *Attributed Relational Graph* foi construído em termos das entidades: instituição, autor (autor e professor), artigo e área de pesquisa. Foram consideradas apenas relações de similaridade entre as entidades de autor. O agrupamento baseado em teoria dos grafos produz as arestas de similaridade apenas nas que possuem uma similaridade mínima (estratégia de esparsificação). A métrica de similaridade tem um valor entre 0 e 1 e tem a conveniência de permitir a composição de similaridade de vários atributos usando proporções.

Observando os dados, ficou evidente que o fato de atributos estarem ausentes em uma instância de um objeto e presentes em outro, como por exemplo, email ou nome completo, não seria bem representado, onde todas as instâncias fossem obrigadas a ser similar entre si (dentro dos critérios de similaridade mínima). Essa situação é descrita em algoritmos de agrupamento hierárquico como a similaridade entre grupos com ligação completa. O outro extremo é a ligação simples, que permite agrupamentos com encadeamento de similaridade, onde há objetos nos extremos muito diferentes entre si. O método Ward fica entre estes extremos, e foi adotado em uma variação bastante eficiente, em termos de tempo de execução, usando o conceito de similaridade estrutural do grafo de dissimilaridade (seção 2.7.2). O grafo de dissimilaridade é construído usando uma matriz de incidência, onde a similaridade entre dois objetos é definida pelo número de arestas diretamente entre eles. A métrica de similaridade entre $[0, 1]$ foi convertida para $[0, 100]$, onde 100 é a similaridade máxima. Este algoritmo está descrito na seção 2.7.2.

Experimentando os agrupamentos deste tipo de dado com as medidas descritas na seção 2.7.4 percebeu-se visualmente que os agrupamentos deveriam ter diâmetro 1 ou até no máximo diâmetro 3. Onde diâmetro de um grupo de elementos, é o maior dentre os caminho mínimo que parte de qualquer par de elementos do grupo no grafo de similaridades que o representa. Desta forma, foi adotado um algoritmo recursivo. Onde, no primeiro passo avalia se componente conexo é muito coeso (se tiver diâmetro 1, ou seja, é um *clique*) ou adotando um critério mais flexível, tendo o diâmetro 3. Um diâmetro

igual a 3 permite que a estrutura do grupo possa conter um *clique*, adicionados de alguns vértices fora do clique que estejam a uma distância máxima de 3 arestas do restante dos outros vértices, sendo assim "coeso o suficiente" para agrupamento de similaridade (esta heurística também permite "cliques" parcialmente completos, também chamados "clubes" ou "comunidades").

Nos componentes conexos que tiverem diâmetro menor ou igual a 3, é considerado um grupo coeso o suficiente, caso contrário aplica-se o algoritmo conforme 2.7.2. Para cada agrupamento obtido no passo anterior que tenha o diâmetro menor ou igual a 3, se aceita como um agrupamento coeso, senão aplica-se o algoritmo recursivamente.

Não foi colocado muito esforço em encontrar o melhor algoritmo de agrupamento para a fase de construção *Attributed Relational Graph*, pois o objetivo é mostrar que as relações encontradas nas redes sociais contidas nestes dados teria influência significativa na melhoria da precisão dos resultados.

Sendo assim, conforme a abordagem de unificação de objetos utilizando seus relacionamentos (seção 2.5), na construção do *Attributed Relational Graph* são consideradas as relações entre as instancias dos objetos: autor e artigo, artigo e instituição, artigo e emails, etc. E conforme descrito nos parágrafos anteriores também são consideradas as relações de similaridade entre os autores: denominações de autores de artigos, denominações de professores.

O próximo passo é o particionamento do *Attributed Relational Graph* baseado na similaridade, no contexto e na força do relacionamento na rede social. O contexto leva em conta informações diretamente relacionadas ao objeto de referência. Este também considera dados encontrados no artigo, onde a denominação do autor foi encontrada (email, instituições, áreas de pesquisa). A rede social escolhida por agregar forte relação no tempo e no espaço entre os participantes foi a rede de coautoria. Desta forma, a sua construção foi realizada de forma a valorizar os indícios de similaridade e também servir como indício de similaridade entre denominações de autores.

A rede social de coautoria e a de autoria, após este processo, permitem o cálculo de medidas estruturais e sumarizações em nível de detalhe e precisão que não está disponível nos dados brutos extraídos da base de dados bibliográfica.

3.2 Extração de informações da base de dados bibliográficos

Nesta seção são descritas as características dos dados da base bibliográfica e na lista de professores de universidades que são entradas da ferramenta, os resultados que devem ser obtidos na saída, e nas etapas do processamento destes dados.

Os dados foram obtidos de consultas na página de internet da Web of Science (seção 2.4), que é uma ferramenta disponível para os pesquisadores identificarem tendências de pesquisa e desempenho de produção científica. Os dados obtidos da consulta devem ser fornecidos ao sistema em arquivos separados por universidades referentes a um intervalo de anos de publicação. No anexo G podem ser encontradas as *strings* de consulta para obter estes dados por universidade.

O formato da base de dados bibliográficos de artigos

Cada registro dos dados da base de dados bibliográfica corresponde a um artigo. Em cada um destes registros há 38 colunas (ou campos) nomeados por uma sigla de 2 letras descritos na Tabela A.1 são listados o significado de algumas destas colunas.

Principais campos do registro de informações bibliográficas de artigos

Alguns dos campos do registro de artigos têm importância para nossa metodologia e têm estrutura complexa, sendo assim merecem uma explicação mais detalhada:

(AU) Autores – lista de nomes de autores separados por ";", os nomes são formatados de forma que o último nome é separado por "," do restante, que é separado por branco.

(AF) Autores abreviados – lista de nomes de autores separados por ";", os nomes são formatados de forma que o último nome é apresentado por extenso e é separado por "," e "branco" do restante, que é uma sequência de caracteres, onde cada caractere é a primeira letra de cada um dos nomes: primeiro nome e os nomes do meio.

(EM) emails – emails dos autores separados por ";". No entanto, não estão relacionados com os respectivos autores.

(C1) Afiliações – frases separadas por ";". No entanto, não estão relacionadas com os respectivos autores. A maioria dos casos respeita o seguinte formato: Grupo Institucional (Instituto, Departamento ou Laboratório), Instituição, Cidade, Estado e País (seção 4.3).

(SC) Áreas de pesquisa – frases padronizadas na WOS que representam categorias de áreas de pesquisa separadas por ";". No apêndice encontra-se a Tabela B.1 que mostra as possíveis frases que representam categorias de áreas de pesquisa utilizadas neste item (SC) (lista de categorias de área de pesquisa).

(UT) Identificação do artigo – esta identificação é única na base, e não há artigos repetidos na base. O WOS utiliza internamente este identificador para relacionar as referências entre os artigos.

(DT) Tipo do documento – esta informação permite que se possam diferenciar dentro todos os tipos de documentos na base os que são artigos e neste caso o valor deste campo é "Article".

Para exemplificar um registro, contendo os campos AU,AF,C1,SC,UT e DT separados pelo caractere de tabulação, tem o seguinte aspecto:
CB Medeiros ;MJ Blin Cláudia Bauser Medeiros IC Unicamp ... Computer Science; Information Systems ISI:000180946400001 Article.

A estes dados são relacionadas categorias de áreas temáticas de pesquisa *Broad-Subject-Fields* e áreas de pesquisa afins, utilizando os valores contidos no campo (SC) dos dados de artigos (Tabela A.1). Com base nos valores deste campo que representam as categorias de pesquisa de periódicos (Tabela B.1) as quais o artigo se refere e as tabelas que relacionam as categorias de pesquisa de periódicos e as áreas de pesquisa **Broad-Subject-Fields** abaixo, cada um dos artigos é classificado em uma ou mais destas:

SCI – Natural Sciences and Mathematics;

ENG – Engineering/Technology and Computer Sciences;

LIFE – Life and Agriculture Sciences;

MED – Clinical Medicine and Pharmacy;

SOC – Social Sciences;

INTER – Interdisciplinary and Multidisciplinary Sciences.

Categorias de áreas de pesquisa afins

Os artigos são classificados também em áreas afins das áreas de pesquisa: *computer science*, *chemistry*, *physics* e *mathematics*, conforme as tabelas indicadas a seguir que as relacionam com categorias de pesquisa de periódicos. Esta informação foi levantada junto a especialistas destas áreas conforme as tabelas abaixo:

CS – Computer Science - Tabela C.1

MATH – Mathematics - Tabela C.2

PHYS – Physics - Tabela C.3

CHEM – Chemistry - Tabela C.4

3.3 O modelo da base de dados

Analisando a base de dados da WOS foram identificados os elementos principais que são entidades e seus atributos importantes para o problema abordado. Nesta seção é apresentada uma visão geral do modelo de dados da ferramenta e seus atributos.

Diagrama da base de dados

No Diagrama 3.1 são apresentadas as principais entidades do problema envolvendo dados bibliográficos.

Nos dados bibliográficos se destacam as denominações de autores, universidades chaves, autores chaves (professores nas universidades chaves), artigos, e relações de autoria.

Na entidade **Artigo**, uma informação importante é o atributo **Broad-Subject-Fields** na entidade **Artigo** contendo a lista de Áreas de Pesquisa de Grande Abrangência **(Broad Subject Fields)**. O atributo **emails** ocorre nas entidades **Artigo** e **TargetProfessor**. No caso da entidade **Artigo** corresponde à lista de emails que estão relacionados a um artigo. No caso da entidade **TargetProfessor** corresponde aos emails dos autores de referência indicados ao sistema. Na seção 4.2 são feitos alguns comentários sobre a importância dos emails para melhorar os resultados de similaridade das denominações de autores e na seção 4.4 é mostrado o uso do atributo **Broad-Subject-Fields** para realizar relatórios consolidados por categorias de áreas de pesquisa.

Na metodologia de agrupamento, destaca-se o Agrupamento dos autores que representa o resultado do tratamento da ambiguidade na denominação dos autores proposta no projeto. O objetivo deste trabalho é realizar análise na rede social que relaciona os agrupamentos de denominações de autores aos seus artigos, produzindo uma estrutura que permite observar a rede coautoria, e a realização de medidas de algumas métricas relacionadas aos autores como a totalização do total de citações e o número de artigos publicados pelo autor baseado nos artigos presentes nesta rede social. Tanto a metodologia como a rede social não estão representados no Diagrama 3.1 que apresenta apenas o modelo dos dados inseridos no sistema.

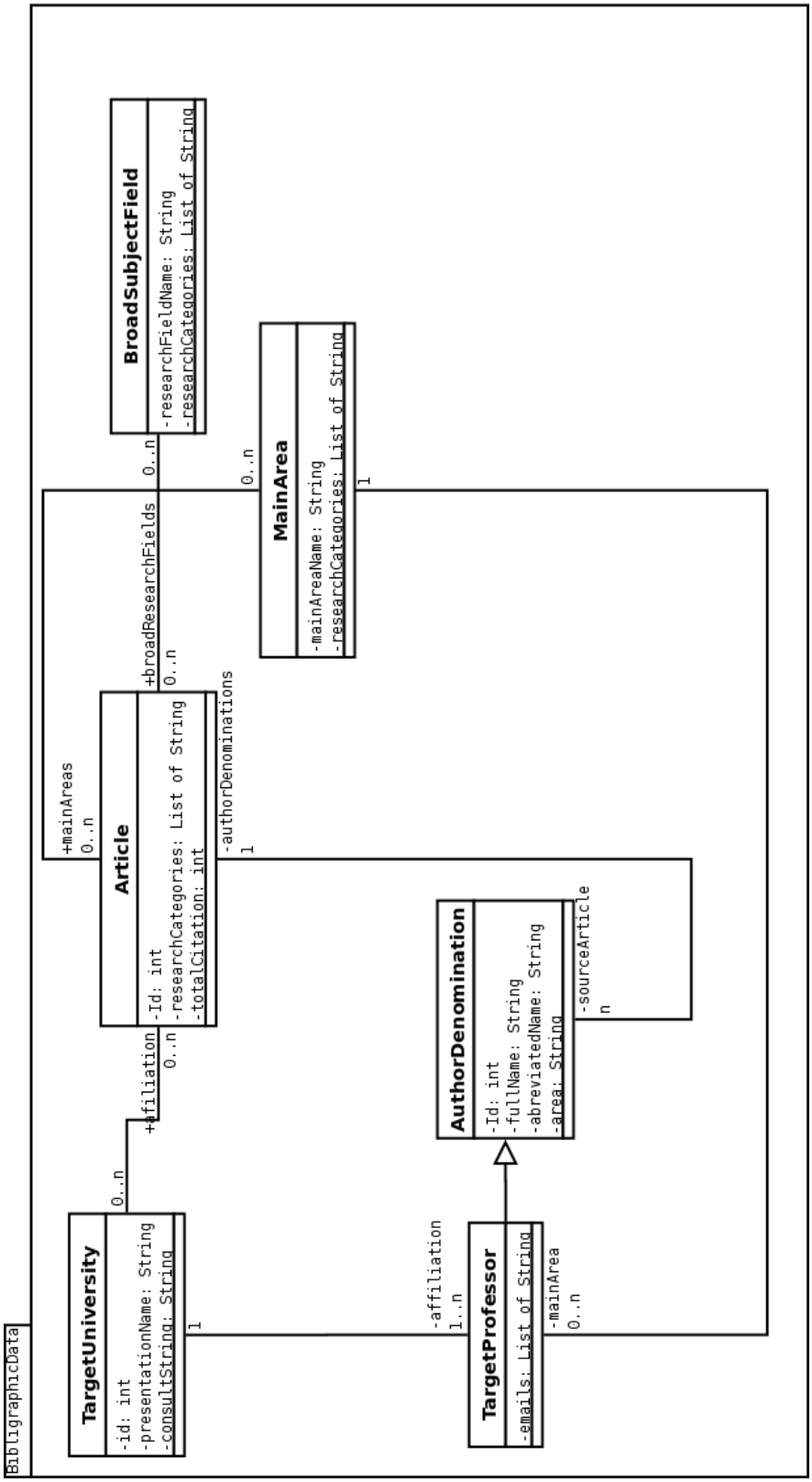

Figura 3.1: Diagrama do modelo da base de dados.

3.4 Projeto e implementação da redução de ambiguidades em nomes

A ferramenta foi implementada usando *Python* [2], que é uma linguagem de programação dinâmica orientada a objetos de propósito geral. Os principais módulos do núcleo básico do *Python* utilizados são *re* (expressões regulares), *files* (manipulação de arquivos), *string* (manipulação de cadeia de caracteres), *shelve* (mecanismo de dicionário hash com persistência de objetos), e alguns módulos adicionais como *scipy* [3] (biblioteca de código aberto e uso livre para pesquisas em aplicações científicas, que oferece um repertório abrangente de operações com matrizes, álgebra linear, álgebra linear esparso, e de análise estatística) e *igraph* [1] (biblioteca de código aberto e de uso livre para criação e manipulação e visualização de grafos, com um bom repertório de algoritmos de teoria dos grafos).

O método de redução de ambiguidades está baseado tanto nos nomes abreviados dos autores, como os nomes completos (informação que nem sempre está disponível nos dados). São considerados também alguns elementos que não são relacionados diretamente com os autores e sim com o artigo, como os emails (que também não estão sempre disponíveis), as universidades que os autores estão afiliados, as categorias de áreas de pesquisa em que se enquadra o artigo e a rede de coautoria que será construída.

3.5 A metodologia

A metodologia para obter os resultados de análise de rede social da base de dados bibliográficos é detalhada nas próximas subseções.

Os cálculos da ferramenta com base nos parâmetros e dados de entrada armazenam alguns dados intermediários que possam ser úteis para outros cálculos que tenham esses mesmos parâmetros e dados de entrada.

A metodologia se divide nas seguintes etapas:

- Definição da área de pesquisa específica que será analisada, para cada universidade do arquivo com a lista de professores de desta área de pesquisa específica (seção 3.5.1)
- Extração dos dados (seção 3.2) e padronização de nomes de autores (seção 3.6.1)
- Redução da Ambiguidade na denominação dos autores e a construção das redes de autoria e de coautoria (seção 3.5.2)
- Cálculo da quantidade de artigos e do total de citações dos artigos de cada universidade por área específica e por Grandes Áreas Temáticas (*Broad Subject Fields*)

(seção 3.5.3). Nesta etapa também é efetuado o cálculo do *ranking* de prestígio das universidades (veja seção 4.5)

- Apresentação dos resultados (seção 3.5.4)

Neste capítulo são detalhados alguns aspectos da metodologia com relação às funções de similaridade:

- Padronização de nomes de autores (seção 3.6.1)
- Funções de similaridade entre nomes de autores (seção 3.6.2)
- Funções de similaridade entre nomes de autores considerando a semântica (vide seção 3.6.3)

3.5.1 Dados de entrada

Os dados de entrada da ferramenta são:

- **O período em anos.**
- **Uma área de pesquisa.**
- **A lista do quadro de professores desta área de pesquisa por universidade**

A lista de professores foi obtida manualmente extraída de páginas de internet das instituições universitárias, com maior renome nas áreas de: Ciência da computação, Matemática, Física e Química dentre as brasileiras e também as de renome mundial. A Lista de professores deve ser fornecida no formato *CSV* (*comma separated value*) contendo os campos: Nome, Nome Abreviado, Universidade, Área, Email, Data.

Na Tabela 3.1 são apresentadas a identificação simplificada e o nome destas universidades. A identificação simplificada dessas universidades é utilizada em várias tabelas no decorrer do texto.

Tabela 3.1: Universidades utilizadas nos experimentos

Identificação simplificada	Nome das universidades
berkeley	University of California, Berkeley
duke	Duke University
mit	Massachusetts Institute of Technology
rice	Rice University
ufmg	Universidade Federal de Minas Gerais
waterloo	University of Waterloo
caltech	California Institute of Technology
harvard	Harvard University
northwestern	Northwestern University
stanford	Stanford University
ufpe	Universidade Federal de Pernambuco
unicamp	Universidade Estadual de Campinas
cmellon	Carnegie Mellon University
illinois	University of Illinois
princeton	Princeton University
syracuse	Syracuse University
ufrgs	Universidade Federal do Rio Grande do Sul
yale	Yale University
cornell	Cornell University
imperialcollege	Imperial College London
puc-rio	Pontifícia Universidade Católica do Rio de Janeiro
texas-austin	University of Texas at Austin
ufrj	Universidade Federal do Rio de Janeiro
usp	Universidade de São Paulo

3.5.2 Redução da Ambiguidade na denominação dos autores

A análise de redes sociais têm como pedra fundamental as medidas estruturais, ou seja, medidas que envolvem relações entre atores do mesmo tipo (autores em relação a autores, artigos em relação a artigos, etc.), ou abstrações de atores (colaboradores na divulgação científica: autores de artigos e veículos de publicação). Além das medidas estruturais existem medidas que envolvem contagens baseadas nas relações de afiliação, que são relações entre atores de classes diferentes (artigos e universidades as quais os autores estão afiliados, artigos e categorias de área de pesquisa). As redes de afiliação a universidades permitem realizar contagens agrupando artigos ou autores por universidades. Lembrando que cada modo em uma rede social corresponde a um tipo de ator participante da rede e no presente projeto correspondem aos artigos, autores, universidades, etc. A grande

maioria das medidas estruturais é aplicável apenas em redes de modo único segundo Wasserman em [45]. Apenas em alguns casos especiais é possível aplicar medidas estruturais em redes de afiliação transformadas em redes de dois modos. Sendo assim, as redes de afiliação também servem para definir agrupamentos de atores em redes de um modo.

Um dos grandes desafios deste projeto é extrair relações estruturais que envolvam atores de mesma classe em dados em que inicialmente as instâncias destes atores não estão estritamente definidas, e algumas das relações entre eles não estão definidas de forma direta. Após as fases do processamento deste componente de redução da ambiguidade na denominação de autores da ferramenta, é possível obter redes sociais que possuem uma variedade maior de análises aplicáveis, como as redes de autoria e as ainda mais interessantes – as redes de coautoria.

O mecanismo de redução de ambiguidade na denominação de autores é composto por quatro fases principais descritas resumidamente a seguir e em mais detalhes nas seções 3.7 e 3.6:

Redes de afiliações - Fase 1: Nesta fase, os dados da base bibliográfica fornecidos ao sistema são identificados e carregados no modelo de dados, juntamente com a lista do quadro de professores de cada universidade relacionada a uma área de pesquisa (autores pivôs). Estes autores pivôs e os artigos obtidos na extração dos dados bibliográficos separados por universidades são identificados unicamente: artigos pelo campo UT e os professores por um identificador numérico único. Desta forma, fica estabelecida a relação de que cada um destes artigos foi produzido por universidades, onde estes artigos constaram no arquivo resultante da consulta aos artigos da base bibliográfica correspondente a esta universidade. Como nas listas de autores pivôs cada um deles também está identificado por universidade estes autores pivôs são afiliados às respectivas universidades.

Observe que se um artigo for produzido por mais de uma universidade ele constará nos dados extraídos para cada uma delas. Assim, fica estabelecida a relação de cada artigo com todas as universidades, onde este artigo conste nos dados que foram fornecidos ao sistema.

Da lista de artigos de todas as universidades, são considerados apenas os artigos que possuam alguma denominação de autor que tenha similaridade como algum autor pivô (seção 4.2). A partir desses parâmetros lista-se a relação de autores (nome, nome abreviado, e-mail, áreas, universidades), onde são padronizados os nomes de autores conforme definido na seção 3.6.1. Esta abordagem, de considerar apenas artigos que tenham autores similares a autores pivôs ou emails de autores pivôs, descarta muitos artigos que não teriam nenhuma relação ou apenas relações remotas com os autores pivôs como é detalhado na seção 4.2.

O modelo de dados descrito na seção 3.3 é carregado, atribuindo a cada denominação de autor (nome abreviado, e nome completo quando presente) em cada artigo e da lista de professores (autor pivô) uma identificação numérica única. Dos dados da lista de artigos são obtidos os identificadores únicos de cada artigo, a lista de identificadores dos autores, as universidades que os produziram e as categorias de pesquisa a que eles se referem.

Estes dados possuem algumas propriedades importantes. Assume-se que:

- um autor pivô está afiliado a uma única instituição acadêmica num dado momento no tempo,
- um autor pivô é definido por um nome completo, um nome abreviado e na maioria dos casos por seus emails,
- os artigos são identificados unicamente pelo campo UT (detalhes na seção 3.2).
- as categorias de pesquisa e os veículos de publicação possuem denominação única.
- cada veículo de publicação se refere a um periódico ou a uma conferência.

Os artigos obtidos da consulta para cada universidade definem, de forma estrita, que estes artigos têm pelo menos um autor afiliado a esta universidade. Mas em geral nesta fase não é possível identificar a qual universidade cada um de seus autores atribuiu a afiliação. Isto porque, nos dados obtidos na base bibliográfica, não existe relação direta entre as denominações de autores e suas afiliações e da mesma forma entre os emails e suas respectivas denominações de autores. Esta rede abriga apenas relações de afiliação, ou seja, a relação de denominações de autores que publicaram um dado artigo, universidades e os artigos publicados por autores afiliados a elas, e assim por diante. Neste tipo de rede social não há relações entre atores de mesma classe, no entanto neste tipo de rede é possível fazer contagens de parâmetros dos atores (vértices) agrupados por estas informações de afiliação (universidades, áreas de pesquisa, periódicos/conferências e áreas temáticas de pesquisa).

Alguns relatórios que a ferramenta produz utilizando esta rede de afiliação realizam sumarizações agrupadas pelas informações de afiliação conforme apresentado na seção 4.4, onde foram consolidadas a produção de artigos por tipo de publicação, por universidades e categorias de pesquisa de todos os artigos carregados no modelo dados da ferramenta.

Rede de autoria - Fase 2 A construção desta rede social de autoria é obtida utilizando a abordagem de "unificação de objetos", onde são calculadas algumas *classes de arestas de similaridade*: entre os nomes de autores completos e nomes abreviados.

Com isto algumas relações regulares também são obtidas: autoria de artigos e afiliação a universidade.

Os experimentos mostram que a informação nesta fase possuiu medidas de qualidade de recuperação de informação de bom nível (valor máximo, como pode ser visto nos resultados apresentados na Tabela 4.7), mas é priorizada a precisão dos dados, que é um pressuposto das análises de redes sociais e este será o enfoque das fases seguintes.

Ao obter a rede de autoria uma relação social importante é explicitada. Nesta rede de autoria é possível identificar a produção científica de cada indivíduo e de grupos de indivíduos quando considerados os autores pivôs (os membros do quadro de professores dos departamentos e institutos de computação conforme dado de entrada ao sistema) como representante do conjunto de denominações agrupadas a ele por similaridade.

Os resultados desta *fase 2* são apresentados na seção 4.2 numa comparação com um algoritmo mais simples e depois comparados com as fases seguintes, em termos de precisão dos resultados na seção 4.6 .

Rede de coautoria inicial - Fase 3: As redes sociais que envolvem apenas uma classe de atores e a relação entre eles são chamadas de redes sociais de um único modo – cada modo de uma rede social corresponde a cada tipo de ator envolvido na sua estrutura. Sendo assim a rede social de autoria possui pelo menos dois modos (artigos e autores), neste tipo de rede fica difícil conceber um significado único para as relações que ocorrem entre os autores (autor-autor) em conjunto com as relações ocorrem entre artigos (artigo-artigo), e entre autores e artigos (autoria). Na realidade a rede de autoria na sua definição comum não há relações (autor-autor) ou (artigo-artigo), a única relação que tem sentido é a de autoria. Por outro lado, conforme pode ser visto na seção 2.3.1 partindo de uma rede de autoria é possível construir uma rede de coautoria correspondente. Desta forma, na rede de coautoria as relações estruturais (autor-autor) têm o significado claro de colaboração científica. No entanto, as relações de afiliação são úteis para consolidar medidas por afiliação a universidades, tipos de publicação, categorias de pesquisa.

Os detalhes da construção da rede de coautoria inicial encontram-se na seção 3.7.1. Para tornar viável a construção desta rede de coautoria foram impostos alguns critérios, exigindo que os participantes da rede possuam aos menos alguns indícios mínimos de que o autor pivô seja o autor do artigo, além disso este procedimento não visa identificar todos os artigos compartilhados entre dois coautores que foram identificados na rede. A informação enumerando os artigos publicados por dois coautores não é fundamental para a rede de coautoria, pois basta que exista um

artigo em coautoria entre dois autores para configurar sua relação de coautoria entre eles.

Rede de coautoria extensa - Fase 4: A rede de coautoria inicial adotou um mecanismo em que identifica as relações de coautoria, mas não objetiva enumerar os artigos envolvidos na relação de coautoria entre dois autores. Nesta *fase 4*, partindo da rede de coautoria inicial, adota-se um novo critério de indício da autoria para uma denominação, que é o de existir um caminho entre esta denominação passando pela rede de coautoria e chegando a outra denominação deste mesmo autor pivô. Todas as denominações de autores com poucos indícios de serem corretamente atribuídas a um autor pivô passam por esta avaliação. As denominações de coautores pertencentes aos agrupamentos de dois autores pivôs, onde uma delas ainda não conste na rede de coautoria, também foram verificadas segundo este critério. Desta forma, este mecanismo trabalha no sentido de completar a informação dos artigos publicados em conjunto entre dois coautores de autores pivôs, ou de reforçar a possibilidade de coautoria, quando uma denominação tem poucos indícios de similaridade com o autor pivô do agrupamento a qual ela faz parte, usando o fato de haver um caminho pela rede de coautoria que as une. As medidas de desempenho na recuperação da informação nos resultados mostraram uma melhora no critério de recuperação priorizando a precisão comparando a *Fase 4* em relação às fases anteriores (na seção 4.6). Com o intuito avaliar a capacidade desta rede de coautoria abstrair a informação de prestígio dos autores, foi criado um *ranking* das universidades baseado no prestígio na rede de coautoria dos autores nelas afiliados. Os resultados foram apresentados na seção 4.5, onde foram comparados com outros *rankings* de universidades da área de engenharia e ciência da computação. Na seção 3.7.2 há mais detalhes sobre a construção da rede de coautoria extensa. A proposta seria repetir a *fase 4* até que nenhuma nova relação de coautoria fosse encontrada, ou que mais artigos fossem identificados nas relações de coautoria. No entanto, não houve tempo para prosseguir o experimento desta forma. A expectativa seria obter uma melhora ainda maior nos resultados apresentados nas seções 4.6 e 4.5.

3.5.3 Sumarização dos resultados

Os artigos representativos nesta consolidação são os artigos escritos pelo quadro de professores em uma área de pesquisa de cada universidade. Esses artigos têm identificadores únicos que nos passos anteriores são obtidas da lista de artigos de cada universidade.

Como existe coautoria entre membros de universidades diferentes e membros da equipe de professores da mesma universidade, este identificador único do artigo permite que

todas as referências a certo artigo sejam consolidadas (foi assumido que esta unicidade é respeitada e que não haja corrupção dos dados).

Cada professor, membro do quadro de professores de cada universidade em uma área específica, irá contribuir com os artigos de sua autoria para a universidade a qual está afiliado, em termos de quantidade de artigos publicados e no total de citações. No entanto, um artigo contribui uma única vez em cada universidade caso tenha sido publicado por mais de um autor afiliado a dada universidade. Os resultados também são agrupados por produção científica por *Broad-Subject-Fields*.

A rede social de coautoria é utilizada para calcular um *ranking* de prestígio das universidades baseado no prestígio dos professores destas universidades nesta rede social conforme detalhado na seção 4.5.

3.5.4 Apresentação dos resultados

Os resultados são apresentados na forma de relatórios:

- com a quantidade de artigos publicados por universidade no período escolhido ou por ano.
- com o total de citações de todos os artigos publicados por universidade no período escolhido ou por ano.

O *ranking* de prestígio das universidades baseado no prestígio de professores destas universidades nesta rede social é apresentado na forma de um relatório contendo o nome da universidade e sua posição. O relatório é ordenado pela posição no *ranking*.

3.6 As funções de similaridade aplicadas na metodologia

3.6.1 Padronização de nomes de autores

Os nomes são padronizados usando sempre letras minúsculas, e colocando primeiro nome, "vírgula", nomes do meio e último nome separados por um branco. Na padronização são substituídas todas as letras acentuadas por letras sem acento ou cedilha. Mantendo após a padronização apenas letras e brancos.

Os nomes abreviados possuem a mesma forma tendo como diferença que o primeiro nome, e nomes do meio são a primeira letra destes seguida de ponto.

3.6.2 Funções de similaridade entre nomes de autores

Neste trabalho foi proposto um mecanismo que consegue capturar a informação de similaridade entre os textos. Esse mecanismo também permite variações na escrita das palavras, utilizando um método geralmente usado no reconhecimento de fala que são os chamados "n-gramas" (um subconjunto de n letras consecutivas de um texto) e que no seu caso mais simples são chamados "bigramas"- que são apenas pares de letras consecutivas presentes no texto [22]. Neste projeto foi colocado o objetivo de garantir que os bigramas contenham o encadeamento das letras, duas a duas e assim pela sua estrutura ele não garante o encadeamento das palavras. Esta abordagem será discutida a seguir, mostrando que abriga algumas vantagens no casamento entre textos que representam nomes que possuam termos abreviados. Quanto ao encadeamento das palavras isto exigiria pelo menos "trigramas", pois o último bigrama de uma palavra será "<última letra><espaço>" e o primeiro será "<espaço><primeira letra>", e se fosse "trigrama" o último "trigrama" de uma palavra seguida da outra seria "<última letra da primeira palavra><espaço><primeira letra da segunda palavra>". No entanto, é interessante não ter a obrigatoriedade do encadeamento de palavras em textos que correspondem aos nomes próprios, como se pode observar por exemplo no caso destas duas denominações válidas da mesma pessoa: Cláudia Maria Bauser Medeiros, ou Cláudia Bauser Medeiros.

A função de similaridade é normalizada de forma a obter um valor entre zero e um e será formalizada no texto abaixo.

Dado que:

$\vec{x}$ e $\vec{y}$: são os vetores binários de presença de bigramas dos elementos X e Y

A função de similaridade é baseada na Similaridade de Cosseno (*Cosine Similarity*) [42]:

$$\begin{aligned} \text{bag of bigrams similarity}(X,Y) &= \text{cosine similarity}(\vec{x},\vec{y}) \\ \text{cosine similarity}(\vec{x},\vec{y}) &= \vec{x}\cdot\vec{y}/(||\vec{x}||\,||\vec{y}||) \\ &= (\vec{x}/||\vec{x}||)\cdot(\vec{y}/||\vec{y}||) \end{aligned}$$

Onde:

$$\begin{aligned} \vec{x}\cdot\vec{y} &= \text{é o produto interno dos vetores } \vec{x} \text{ e } \vec{y}. \\ ||\vec{x}|| &= \sqrt{(\vec{x}.\vec{x})} \text{ ou o comprimento do vetor } \vec{x}. \end{aligned}$$

Ou seja, segundo as formulas acima se as denominações sendo comparadas forem iguais o valor da similaridade será um (1), e se forem diferentes será zero (0).

Nesta fórmula é assumido que:

$$\text{bag of bigrams proximity} \quad = \quad 1 - \text{bag of bi-grams similarity}$$

A Similaridade de Cosseno é geralmente usada para vetores de palavras referentes a um documento, mas neste caso são vetores de bigramas de um nome. A Similaridade de Cosseno respeita a desigualdade triangular, sendo assim em trabalhos futuros esta propriedade poderá ser usada para reduzir o número de comparações no algoritmo de identificação de componentes conexos do grafo de similaridade.

3.6.3 Funções de similaridade entre nomes de autores usando semântica

A função de *bag of bigrams similarity* é composta com outra função denominada *compare authors* que considera os significados de algumas palavras e também a estrutura básica dos nomes de autores. A função *compare authors* incorpora os seguintes conceitos:

- descarta nomes cujas primeiras letras são totalmente diferentes (usando função de similaridade por string kernel descrita na seção 2.6.2, implementada pelo projeto [15])
- prioriza a versão por extenso do nome,
- descarta ponto(.), traço (−),sinais(+),
- descarta também algumas palavras: e, de, dad, do, do, das, dos e padroniza: *jr* em *junior*.
- garante igualdade entre primeiro nome, e ultimo nome,
- e garante que as combinações dos nomes intermediários possam casar entre nomes por extenso e a versão abreviada no outro nome. O casamento permite omissões de nomes intermediários desde que os próximos casamentos mantenham a ordem.

Dentre as combinações possíveis calculadas é escolhida a que tenha maior casamento em número de caracteres e então o valor do maior casamento é normalizado pelo tamanho do maior nome (em caracteres) dentre os dois sendo comparados.

A função composta das funções de similaridade *bag of bigrams similarity* e a *compare authors* é o resultado da média geométrica entre as duas funções de similaridade (média ponderada *bag of bigrams similarity* de nomes abreviados e nomes por extenso) com a *compare authors*. A média geométrica foi escolhida, pois valoriza a identificação de similaridade por cada um dos participantes da média.

Houve grandes desafios em termos de volume de dados, como pode ser visto nas tabelas que mostram o tamanho de nossa base de dados por universidades e áreas de pesquisa na seção 4.4.

A solução tomada foi considerar apenas os artigos de uma dada área de pesquisa, e os homônimos de professores dentre os artigos publicados nas suas universidades, como será apresentado mais detalhes na discussão dos resultados na seção 4.1.

Nos ensaios foi percebido que a distância entre nomes de autores necessitava de uma parcela de semântica da estrutura de nomes próprios.

A função *compare authors* incorpora conceitos que levam em conta o significado das palavras e a estrutura dos nomes próprios de pessoas foi utilizada para diminuir o volume de dados a ser processado.

A função baseada em *bag of bigrams* em relação à similaridade se apresentou mais forte e mais rápida. Sendo assim foi feita uma otimização: ao atingir dissimilaridade acima de *1-similaridade mínima* nem se realiza o cálculo da *compare authors*, que é bem mais custosa em termos de tempo de execução. Exemplos concretos podem ser encontrados na seção 4.1.

3.7 A construção da rede social de coautoria

A construção da rede social de coautoria é apresentada brevemente nesta seção em duas fases descritas com mais detalhes nas próximas seções.

3.7.1 A construção da rede social de coautoria inicial

A construção de uma rede de coautoria numa situação ideal exigiria que houvesse uma relação estrita entre autores pivôs (o quadro de professores das universidades fornecidos como dados de entrada à ferramenta), seus coautores (identificados unicamente) e os artigos que publicaram. No caso dos dados que a ferramenta utiliza há apenas prováveis denominações que correspondam aos autores pivôs, e denominações pouco ou nada similares que correspondem aos seus coautores (seção 3.6.2) e esta é a informação que é produzida na construção da rede de autoria. A rede de autoria segrega denominações semelhantes a um autor pivô cuja afiliação a universidades dos artigos, onde são encontradas, caso não possuam a universidade deste autor pivô. O pressuposto de que a universidade foi

corretamente identificada na consulta à base de dados bibliográfica permite utilizar uma heurística. Que dentre os autores pivôs, o autor pivô afiliado a uma universidade mais similar aos autores de artigos atribuídos a esta universidade é a denominação de autor pivô com maior indício de ser a correta a atribuir a estes autores.

Como ponto inicial este algoritmo, irá unificar prováveis denominações de autores pivôs a cada autor pivô de forma mais exigente e assim, conceder mais precisão à rede de autoria e por conseguinte às relações da rede de coautoria inicial. Esta abordagem permite que se possa considerar os dados mais relevantes para uma rede de coautoria que envolva os autores pivôs. Sendo assim, foram descartados os componentes da rede de coautoria que não envolviam os autores-pivôs. Para isso foram estabelecidos alguns critérios que buscam encontrar dentre as prováveis denominações de um autor pivô, ou seja, as denominações dos autores de cada artigo da base de dados que possuam indícios mínimos de corresponderem a este autor pivô. A função *Indícios de similaridade suficientes*, descrita no algoritmo 3, levam em consideração dados constantes no artigo na qual se encontra aquela denominação: emails, categorias de área de pesquisa, e a sua similaridade com autor pivô de seu agrupamento obtido na construção da rede de autoria. Ou seja, similaridade entre denominações, se ele conta com o nome completo além do abreviado, se o email o autor pivô é encontrado na lista de emails do artigo e se área de pesquisa é a mesma da definida para o autor pivô.

A construção da rede de coautoria inicial se dá em dois passos: o primeiro considera apenas os agrupamentos de autores pivôs e os artigos em comum entre eles, o segundo passo considera os artigos diferentes entre os agrupamentos de autores pivôs, mas que tenham algum coautor em comum (que não é um agrupamento de autor pivô). Por isso é chamada de rede de coautoria inicial. O produto deste algoritmo é uma matriz que identifica a possibilidade de coautoria entre dois agrupamentos de denominações de autores, e a lista de denominações que foram identificadas nesta rede de coautoria.

Função de decisão se há indícios de similaridade suficiente entre uma denominação e um autor pivô

A construção da rede de coautoria utiliza a função *Indícios de similaridade suficientes* (algoritmo 3) que decide se há indícios de similaridade suficiente entre uma denominação e um autor pivô. Esta função tem os seguintes parâmetros: Indícios de similaridade entre denominação de autor de artigo e uma denominação de autor pivô e o *SimMin* ou Limiar de mínima similaridade entre denominações.

Indícios de similaridade entre uma denominação de autor e um autor pivô identificam situações chamadas de indícios de similaridade: emails em comum (email na lista de emails do artigo e os emails do autor pivô), domínio de email em comum (o mesmo em relação ao domínio do email), área de pesquisa em comum, área temática de pesquisa em comum

Algoritmo 1: Contrutor da rede de coautoria inicial - passo 1

Entrada: IdsAgrupamentosPivos, AgrupamentosDenAutores, SimMin
Saída: MatrizPossibilidadeCoautoria, DenNaRede

Incializa com zeros MatrizPossibilidadeCoautoria e DenNaRede=$\emptyset$;
para todo IdAgX, IdAgZ $\in$ IdsAgrupamentosPivos ***e*** IdAgX $\geqq$ IdAgZ **faça**
  AgX,AgZ = agrupamentos em AgrupamentosDenAutores com IdAgX e de IdAgZ;
  AutorPivoX,AutorPivoZ = denominações dos autores pivôs do AgX e de IdAgZ ;
  EmailsPX = emails de AutorPivoX;
  EmailsPZ = emails de AutorPivoZ;
  para todo DenAutorX $\in$ AgX ***e*** DenAutorZ $\in$ AgZ **faça**
    ArtigoX = artigo a que se refere DenAutorX;
    ArtigoZ = artigo a que se refere DenAutorZ;
    EmailsAX = emails do ArtigoX;
    EmailsAZ = emails do ArtigoZ;
    s_x = similaridade entre AutorPivoX e DenAutorX;
    s_z = similaridade entre AutorPivoZ e DenAutorZ;
    $s_x = s_x$ **ou** 100 **se** EmailsPX $\cap$ EmailsAX $\neq \emptyset$;
    $s_z = s_z$ **ou** 100 **se** EmailsPZ $\cap$ EmailsAZ $\neq \emptyset$;
    IndSimXPX = indícios de similaridade entre AutorPivoX e DenAutorX;
    IndSimZPZ = indícios de similaridade entre AutorPivoZ e DenAutorZ;
    se ArtigoX = ArtigoZ ***e*** $s_x \geqq$ SimMin ***e*** $s_z \geqq$ SimMin **então**
      se *IndiciosSimilaridadeSuficientes(s_x,*IndSimXPX*)* ***ou***
      *IndiciosSimilaridadeSuficientes(s_z,*IndSimZPZ*)* **então**
        MatrizPossibilidadeCoautoria[IdAgX, IdAgZ] = $(s_x s_z)/100$;
        DenNaRede = DenNaRede $\cup$ {DenAutorX, DenAutorZ};
      fim
    fim
  fim
fim

Algoritmo 2: Contrutor da rede de coautoria inicial - passo 2

Entrada: IdsAgrupamentosPivos, AgrupamentosDenAutores, SimMin
Saída: MatrizPossibilidadeCoautoria, DenNaRede

para todo IdAgX, IdAgZ ∈ IdsAgrupamentosPivos ***e*** IdAgX ≧ IdAgZ **faça**
  AgX,AgZ = agrupamentos em AgrupamentosDenAutores com IdAgX e de IdAgZ;
  AutorPivoX,AutorPivoZ = denominações dos autores pivôs do AgX e de IdAgZ ;
  EmailsPX,EmailsPZ = emails de AutorPivoX e de AutorPivoZ;
  para todo DenAutorX ∈ AgX ***e*** DenAutorZ ∈ AgZ **faça**
    ArtigoX = artigo a que se refere DenAutorX;
    ArtigoZ = artigo a que se refere DenAutorZ;
    EmailsAX,DenXs = emails e autores em artigo ArtigoX;
    EmailsAZ,DenZs = emails e autores em artigo ArtigoZ;
    IdAgYXs = identificadores dos agrupamentos dos DenXs diferentes de IdAgX;
    IdAgYZs = identificadores dos agrupamentos dos DenZs diferentes de IdAgZ;
    IdAgYs = IdAgYXs ∩ IdAgYZs;
    para todo IdAgY ∈ IdAgYs ***e*** ArtigoX ≠ ArtigoZ **faça**
      AgY = agrupamento em AgrupamentosDenAutores com IdAgY;
      s_x = similaridade entre AutorPivoX e DenAutorX;
      s_z = similaridade entre AutorPivoZ e DenAutorZ;
      $s_x = s_x$ **ou** 100 **se** EmailsPX ∩ EmailsAX ≠ ∅ ;
      $s_z = s_z$ **ou** 100 **se** EmailsPZ ∩ EmailsAZ ≠ ∅ ;
      IndSimXPX = indícios de similaridade entre AutorPivoX e DenAutorX;
      IndSimZPZ = indícios de similaridade entre AutorPivoZ e DenAutorZ;
      para todo (DenAutorY, DenAutorYb) ∈ AgY ***e*** DenAutorY ∈ DenXs ***e***
      DenAutorYb ∈ DenZs **faça**
        s_{y,y_b} = similaridade entre DenAutorY e DenAutorYb;
        $s_{x,y} = (s_x s_{y,y_b})/100$;
        $s_{z,y} = (s_z s_{y,y_b})/100$;
        $s_{x,z} = (s_{x,y} s_{z,y})/100$;
        se $s_{x,z}$ ≧ SimMin ***e*** *(IndiciosSimilaridadeSuficientes(s_x,*IndSimXPX*)*
        ou *IndiciosSimilaridadeSuficientes(s_z,*IndSimZPZ*))* **então**
          atualiza MatrizPossibilidadeCoautoria com $s_{x,z}$, $s_{z,y}$, $s_{z,y}$ desde que já não haja um valor maior;
          DenNaRede = DenNaRede ∪ {DenAutorY, DenAutorYb};
      fim
    fim
  fim
fim

presença de nome completo na denominação de autor.

O limiar de mínima similaridade (*SimMin*) é um parâmetro da ferramenta como um todo e define em um significado numérico a exigência de similaridade mínima em três patamares: estreito, muito estreito e extremamente estreito. Estes patamares de similaridade podem ser otimizados em trabalhos futuros, utilizando algoritmos de aprendizado baseados em exemplos com uma amostra de dados previamente catalogada.

Algoritmo 3: IndíciosSimilaridadeSuficientes

Entrada: SimilaridadeEntreDenominaçãoAutorEDenominaçãoAutorPivô, IndíciosSimilaridadeEntreDenominaçãoAutorEDenominaçãoAutorPivô
Saída: verdadeiro ou falso

```
SimXPX = SimilaridadeEntreDenominaçãoAutorEDenominaçãoAutorPivô;
IndXPX = IndíciosSimilaridadeEntreDenominaçãoAutorEDenominaçãoAutorPivô;
SimMinEstreito = 100 − (100 − SimMin)/2;
SimMinMuitoEstreito = 100 − (100 − SimMin)/4;
SimMinExtremamenteEstreito = 100 − (100 − SimMin)/8;
selecione IndXPX faça
    caso IndXPX indica email em comum com autor pivô
        retorne verdadeiro
    caso IndXPX indica área de pesquisa em comum com autor pivô
        se SimXPX ≧ SimMinEstreito e IndXPX indica que denominação tem
        especificado seu nome completo então
            retorne verdadeiro
        senão
            se SimXPX ≧ SimMinMuitoEstreito então
                retorne verdadeiro
            senão
                retorne falso
            fim
        fim
    caso IndXPX indica que denominação tem especificado seu nome completo
        se SimXPX ≧ SimMinExtremamenteEstreito então
            retorne verdadeiro
        senão
            retorne falso
        fim
    senão retorne falso
fim
```

3.7.2 A construção da rede social de coautoria extensa

A rede de coautoria inicial adota um mecanismo que identifica as relações de coautoria, mas não objetiva à enumeração dos artigos envolvidos na relação de coautoria entre dois autores. Nesta etapa, partindo da rede de coautoria inicial, adota-se um novo critério de indício da autoria para uma denominação, que é o de existir um caminho entre esta denominação passando pela rede de coautoria e chegando a outra denominação deste mesmo autor pivô.

Todas as denominações de autores com poucos indícios de serem corretamente atribuídas a um autor pivô passam por esta avaliação. As denominações de coautores pertencentes aos agrupamentos de dois autores pivôs, onde uma delas ainda não conste na rede de coautoria, foram verificadas segundo este critério. Desta forma, este mecanismo trabalha no sentido de completar a informação dos artigos publicados em conjunto entre dois coautores de autores pivôs, ou de reforçar a possibilidade de coautoria, quando uma denominação tem poucos indícios de similaridade com o autor pivô do agrupamento a qual ela faz parte, usando o fato de haver um caminho pela rede de coautoria que as une. As denominações de autores, que são o ponto de partida para encontrar o caminho pela rede de coautoria, são obtidas através da função *CandidatosPartida* e está no algoritmo 6.

Essas denominações de autores candidatas de partida têm poucos indícios de serem corretamente atribuídas a um autor pivô, ou fazem parte de pares autores pivôs que tenham artigos que ainda não foram inseridos na rede de coautoria, denominados de candidatos a ponto de partida do caminho na rede de coautoria. Esta função apesar de exigir quase nada da denominação com poucos indícios de serem corretamente atribuídas a um autor pivô, estabelece a exigência de que a primeira relação de coautoria seja com uma denominação de autor pertencente a um agrupamento com autor pivô, ou que ambas pertençam a agrupamentos de autores pivôs.

A medida de similaridade mínima exigida será ponderada pelo critério de mínima similaridade estabelecido para o caminho na rede social, que neste aspecto estará medindo a mínima possibilidade de que as relações de coautoria sejam corretamente representadas pelo parâmetro *SimMinCaminhada* do algoritmo 4, que descreve construção da rede de coautoria extensa.

Os possíveis pontos de chegada deste caminho pela rede de coautoria são obtidos pela função *ObterCandidatosChegada* (algoritmo 5). Esta função estabelece critérios mais rigorosos para o ponto de chegada de volta ao agrupamento, cujos elementos estão sendo ponto de partida para o caminho, para que o caminho sirva de indício suficiente para considerar alguns destes candidatos de partida como relacionados ao autor pivô deste agrupamento, mesmo esta denominação de autor, ela mesma, tendo poucos indícios disto.

O critério adotado para o ponto de chegada exige indícios de similaridade usando a função *Indícios de similaridade suficientes* (vide algoritmo 3) tanto para o ponto de che-

gada ao agrupamento (denominação pertencente ao agrupamento), quanto para a denominação em algum artigo que esta denominação é coautora. Este coautor deverá pertencer a um agrupamento com autor pivô e também possuir indícios de similaridade mínimos (algoritmo 3) com este autor pivô.

Finalmente é descrito o algoritmo que percorre a rede de coautoria dos pontos de partida e busca encontrar os pontos de chegada. Este algoritmo, chamado de *CaminhadaPelaRedeDeCoautoria* (algoritmo 7), utiliza conceitos semelhantes ao do passeio aleatório em grafos cujas arestas representam probabilidades de passar de um vértice para outro. No entanto, no caso deste algoritmo as arestas desta rede de coautoria, que como a própria matriz que a define denota *MatrizPossibilidadeCoautoria*, representa apenas a possibilidade que haja coautoria entre dois autores pivôs, ou entre agrupamentos de denominações semelhantes. Ou seja, ao passar de um autor para outro via a relação de coautoria a possibilidade de coautoria até aquele ponto deve ser ponderada pela obtida no passo anterior e assim por diante.

Um tamanho máximo foi estabelecido como corte da caminhada, que é representado pelo parâmetro *CaminhoMax*. Por fim, quando encontra um caminho entre estas duas denominações, que diretamente têm poucos indícios de corresponderem ao mesmo autor pivô, então, a possibilidade ponderada encontrada no caminho é atribuída à relação de coautoria entre este autor pivô e o outro que participou da relação de coautoria de partida do caminho (a possibilidade de coautoria ponderada do caminho dada por *PossibilidadeCoautoriaXY*).

O valor assumido para a relação entre *SimMinCaminha* e *SimMin* é dada por:

$$\frac{SimMinCaminha}{SimMin} = 0,24 \ ,$$

Esta relação pode ser ajustada futuramente com técnicas de aprendizado. O tamanho máximo do caminho ficou fixo em 6 e este também pode ser avaliado em futuras otimizações.

No apêndice H são apresentados exemplos do funcionamento deste algoritmo e no próximo capítulo são apresentados os resultados.

Algoritmo 4: Construtor da rede de coautoria extensa

Entrada: IdsAgrupamentosPivos, AgrupamentosDenAutores, SimMinCaminhada, CaminhoMax

Saída: MatrizPossibilidadeCoautoria, DenNaRede

Inicializa MatrizPossibilidadeCoautoriaAux com zeros e DenNaRedeAux $= \emptyset$;
para todo IdAgX $\in$ IdsAgrupamentosPivos **faça**
  AgX = agrupamentos em AgrupamentosDenAutores com IdAgX;
  CandidatosChegadaX = `ObterCandidatosChegada(IdAgX)` ;
  CandidatosPartidaX = `ObterCandidatosPartida(IdAgX)` ;
  para todo (DenAutorX, DenAutorY) $\in$ CandidatosPartidaX ***e***
  CandidatosChegadaX $\neq \emptyset$ **faça**
    (EncontradoCaminho,PossibilidadeCoautoriaXY) =
    `CaminhadaPelaRedeDeCoautoria(` DenAutorY, IdAgX, CandidatosChegadaX, SimMinCaminhada, CaminhoMax, MatrizPossibilidadeCoautoria`)` ;
    se EncontradoCaminho **então**
      IdAgY = identificador do agrupamento de DenAutorY;
      $s_{x,y}$ = MatrizPossibilidadeCoautoria[IdAgX, IdAgY] ;
      $s_{x,y}$ = Máximo de $\{s_{x,y}$, PossibilidadeCoautoriaXY$\}$;
      MatrizPossibilidadeCoautoriaAux[IdAgX, IdAgY] $= s_{x,y}$;
      DenNaRedeAux = DenNaRedeAux $\cup$ {DenAutorX, DenAutorY} ;
  fim
fim

DenNaRede = DenNaRede $\cup$ DenNaRedeAux ;
MatrizPossibilidadeCoautoria = `Maximum(` MatrizPossibilidadeCoautoriaAux, MatrizPossibilidadeCoautoria`)` ;

Algoritmo 5: ObterCandidatosChegada

Entrada: IdAg, IdsAgrupamentosPivos, AgrupamentosDenAutores, SimMinCaminhada
Saída: `CandidatosChegada`

```
Ag = agrupamento em AgrupamentosDenAutores com IdAg;
CandidatosChegada = ∅ ;
AutorPivo = autor pivô do agrupamento Ag;
para todo DenAutor ∈ Ag e DenAutor ∈ DenNaRede faça
    s_x = similaridade entre AutorPivo e DenAutor;
    IndSimXPX = indícios de similaridade entre AutorPivo e DenAutor;
    se
    IndiciosSimilaridadeSuficientes(s_x,IndSimXPX(SimMinCaminhada/SimMin))
    então
        Artigo = artigo a que corresponde DenAutor;
        DenAutorZs = denominações dos autores de Artigo;
        para todo DenAutorZ ∈ DenAutorZs e DenAutorZ ≠ DenAutor faça
            AutorPivoZ = autor pivô do agrupamento de DenAutorZ;
            s_z = similaridade entre AutorPivoZ e DenAutorZ;
            IndSimZPZ = indícios de similaridade entre AutorPivoZ e DenAutorZ;
            se
            IndiciosSimilaridadeSuficientes(s_z,IndSimZPZ(SimMinCaminhada/SimMin))
            então
                CandidatosChegada = CandidatosChegada ∪ (DenAutorZ) ;
            fim
        fim
    fim
fim
```

Algoritmo 6: ObterCandidatosPartida

Entrada: IdAg, IdsAgrupamentosPivos, AgrupamentosDenAutores, SimMinCaminhada
Saída: `CandidatosPartida`

Ag = agrupamento em AgrupamentosDenAutores com IdAg;
`CandidatosPartida` = $\emptyset$;
AutorPivo = autor pivô do agrupamento Ag;
para todo DenAutor $\in$ Ag ***e*** DenAutor $\in$ DenNaRede **faça**
  Artigo = artigo a que corresponde DenAutor;
  DenAutorZs = denominações dos autores de Artigo;
  para todo DenAutorZ $\in$ DenAutorZs ***e*** DenAutorZ $\neq$ DenAutor **faça**
    IdAgZ = identificador do Agrupamento ao qual pertence DenAutorZ;
    `// por definição DenAutor é uma`
    `// denominação que pertence a um agrupamento de autor pivô`
    `// ela está sendo cogitada para ponto de partida`
    `// e é aceita se:`
    `// 1) não pertence a rede de coautoria e tem`
    `// coautoria com outra de um agrupamento de autor pivô`
    `// 2) está na rede de coautoria e a do coautor não`
    se *(* DenAutor ***não*** $\in$ DenNaRede ***e*** IdAgZ $\in$ IdsAgrupamentosPivos *)* ***ou*** *(* DenAutor $\in$ DenNaRede ***e*** DenAutorZ ***não*** $\in$ DenNaRede *)* **então**
      `CandidatosPartida` = `CandidatosPartida` $\cup$ (DenAutor, DenAutorZ) ;
    fim
  fim
fim

Algoritmo 7: CaminhadaPelaRedeDeCoautoria

Entrada: DenAutorY, IdAgC, CandidatosChegada, SimMinCaminhada, CaminhoMax, MatrizPossibilidadeCoautoria
Saída: EncontradoCaminho,PossibilidadeCoautoriaXZ

IdAgP = índice do agrupamento a qual pertence DenAutorY;
IdAgC = IdAgC;
IdAgsCandidatosChegada = índices dos agrupamentos da qual fazem parte as denominações de autores em CandidatosChegada v_0 = vetor do tamanho de uma dimensão de MatrizPossibilidadeCoautoria;
v_0[IdAgP] = 1 ;
$v_1 = v_0$;
Caminho =0 ;
EncontradoCaminho = falso;
PossibilidadeCoautoriaXZ = 0 ;
PossibilidadeMuitoBaixa = falso;
IdsAgrupamentosAdicionados = $\emptyset$ **enquanto** ***não*** PossibilidadeMuitoBaixa ***e não*** EncontradoCaminho ***e*** Caminho $++ \leqq$ CaminhoMax $- 1$ **faça**
    `// produto vetorial entre` v_1 `e` MatrizPossibilidadeCoautoria
    M=v_1 . MatrizPossibilidadeCoautoria;
    M=M/100 ;
    $M = \{\ M_{i,j} = 1$ **se** $M_{i,j} \geqq$ SimMinCaminhada **ou** 0 **caso contrário** $\}$;
    $v_1 = \{\ v_{1_i} =$ `Minimum(`$M_{i,j}$`,100)` para todo $j\ \}$;
    IdsAgrupamentosNoPasso = $\{\ i$ se $v_{1_i} \neq 0\ \}$;
    IdsAgrupamentosAdicionados =
    IdsAgrupamentosAdicionados $\cup$ IdsAgrupamentosNoPasso ;
    se $v_1[IdAgC] \geqq$ SimMinCaminhada **então**
        `// Possível caminho entre` IdAgP `e` IdAgC
        se IdAgC $\in$ IdsAgrupamentosAdicionados ***e***
        IdAgsCandidatosChegada $\cap$ IdsAgrupamentosAdicionados $\neq \emptyset$ **então**
            EncontradoCaminho = verdadeiro;
            PossibilidadeCoautoriaXZ = v_1[IdAgC] ;
        se `Maximum(` v_1`)` $== 0$ **então**
            `// Possibilidade muito baixa de haver caminho entre` IdAgP `e` IdAgC
            PossibilidadeMuitoBaixa = verdadeiro;
fim
retorna (EncontradoCaminho, PossibilidadeCoautoriaXZ) ;

Capítulo 4

Resultados

Neste capítulo são destacados os resultados obtidos aplicando a metodologia proposta para viabilizar a análise estrutural em redes sociais de colaboração científica a partir de bases de dados bibliográficos.

A máquina utilizada para os experimentos foi uma Dell Poweredge T300 com processador Quad Core Xeon X3363, com processador 2x6M de Cache, 2,83 GHz, 1333MHz de FSB e 3Gb de RAM, comprada com a verba técnica da Fapesp, onde foi possível executar uma versão do experimento em 2 dias para uma dada área de pesquisa.

4.1 Melhorias na metodologia utilizando semântica

Um dos grandes desafios para o mecanismo de redução da ambiguidade na denominação dos autores foi o volume de dados, como pode ser visto nas tabelas que mostram o tamanho de nossa base de dados por universidades e áreas de pesquisa. O volume total de artigos na base analisada é de 291.375, levando a uma variedade de denominações diferentes para autores em torno de 1 milhão.

Como solução para reduzir o volume de dados utilizados foram feitas algumas restrições em relação à quais artigos considerar:

- os artigos de uma dada área ou;
- os artigos de homônimos de professores (lista de autores referência) ou;
- artigos de autores com denominações similares as dos nomes dos professores (lista de autores referência) dentre a totalidade de artigos publicados nas universidades destes professores.

Esta abordagem resultou em um volume de dados tratável, ou seja, em torno de dezenas de milhares de artigos, com uma variedade de denominações de autores em torno

de menos de meia centena de milhares.

A razão intuitiva desta escolha é que o grau de incerteza aumenta muito quando envolve denominações de autores de artigos que não são similares a professores (autores pivôs) e não são da mesma área de pesquisa destes professores mesmo que eles pertençam à mesma universidade.

O problema prático abordado nos experimentos é o de realizar a análise de redes sociais relacionadas a universidades em uma dada área de pesquisa, com base na produção científica dos professores do departamento (ou instituto) - que representam esta área de pesquisa na instituição. Essas restrições não afetaram este objetivo como pode ser observado nos resultados apresentados neste capítulo.

A abordagem de prévia seleção dos dados a serem processados pelo mecanismo de sumarização dentre o todo dos dados de entrada é utilizada em problemas semelhantes que sucumbem frente ao volume de dados [42].

Outro argumento que segue nesta direção é que o grafo de coautoria apresentou-se muito largo, o diâmetro era da ordem de 30, e daí por diante as similaridades entre as denominações de autores ficam mais e mais remotas. Desta forma para a melhoria do grau de confiança nos resultados foi necessário agregar mais atributos que o nome, como afiliação do autor e o provável email. Isto se faz mais necessário quando a denominação do autor não tem semelhança com um dos nomes da lista do quadro de professores das instituições. Na seção seguinte são detalhados os experimentos feitos aplicando esta abordagem de pré-processamento para diminuir o volume cálculos de similaridade usando nomes de autores e emails.

4.2 Redução do volume cálculo de similaridade usando nomes de autores e emails

Alguns experimentos foram realizados com a função de similaridade *compare authors* (seção 3.6.3) para garantir que, na fase final do agrupamento, os nomes dos agrupamentos respeitassem alguns parâmetros de formação de nomes que esta função propõe.

Como *compare authors* é estrita em algumas decisões, esta acaba descartando alguns nomes que tiveram erros de digitação. Então com a hipótese de que a função de similaridade de *bag of bigrams de nomes completos e abreviados*, fosse mais permissiva do que a *compare authors*, colocou-se a função de similaridade *bag of bigrams de nomes completos e abreviados* na primeira fase do mecanismo de agrupamento (identificação de componentes conexos no grafo de similaridade) que é a mais pesada em termos de volume de dados, e no tempo de processamento (onde o maior custo se concentra em obter os componentes conexos).

A função de similaridade *compare authors* tem um tempo de execução mais custoso, apesar de mais econômica em termos de memória. O resultado desta função se mostrou bem melhor, apesar de uns poucos descartes de nomes com erros de digitação.

No entanto, ao comparar com um método que usa apenas a função de similaridade *compare authors* e calcula apenas a relação de similaridade dos nomes de autores de artigos de uma dada universidade e o grupo de professores referência, percebeu-se que a função similaridade de *bag of bigrams de nomes completos e abreviados* estava descartando alguns poucos nomes de autores, obtidos neste mecanismo mais simples, e casando com um número muito maior de variações de denominações de autores. Mas, autores muito ativos como *Berthier Ribeiro-Neto* estavam sendo descartados. Este mecanismo mais simples é denominado BIPARTITE, por ter uma abordagem que leva uma única aresta partindo de cada um dos elementos a serem agrupados e que incide sobre apenas um dos elementos que são referência ou protótipos de denominação de autores (lista do quadro de professores dos departamentos ou institutos das universidades).

Sendo assim, o grande desafio que se colocou foi melhorar a seleção inicial dos dados que iriam ser processados no algoritmo de agrupamento. Desta forma, foram considerados todos os atributos que valorizam a similaridade entre autores e entre autores e colegas de "departamento", e para isso os e-emails do quadro de professores, que foram obtidos das páginas das universidades manualmente para Computação e Química.

Esta abordagem melhorou bem o resultado no sentido de acrescentar aos artigos anteriormente selecionados, os artigos que são de outras áreas na mesma universidade ou artigos da mesma área estudada, mas cujos nomes tinham grande diferença em termos de quantidade de letras, no entanto considerando os fatores semânticos que a função de similaridade *compare authors* considera, obteve-se uma similaridade mais pertinente, permitido o agrupamento destas denominações com os devidos autores pertencentes ao grupo de professores selecionados como entrada do processo.

Seguindo desta forma obtiveram-se resultados concretos em relação à unificação de nomes de autores. Os resultados dos agrupamentos feitos com o algoritmo deste projeto (que se chamou de *AREA*) foram comparados com outro chamado de *BIPARTITE* que é detalhado a seguir. O algoritmo *BIPARTITE* foi assim chamado por usar uma metodologia de bipartição (*BIPARTITE*), que pode ser modelado como um grafo em que os vértices são de dois tipos: de um lado ficam os autores de artigos de uma universidade e de outro os nomes dos professores selecionados, e entre os dois lados calcula-se o melhor casamento entre cada um dos autores com o quadro de professores, com base na similaridade entre autores e professores. O cálculo da similaridade para o *BIPARTITE* foi feito usando apenas a função de similaridade *compare author*.

Em resumo o resultado desta comparação para os dados da área de Ciência da Computação (CS) foi:

- o algoritmo BIPARTITE encontrou 131 artigos para o quadro de professores do Instituto de computação da Unicamp e o algoritmo AREA encontrou 199 (dentre os agrupamentos de professores 4 tem elementos discutíveis) (seção 4.2),
- o algoritmo BIPARTITE encontrou 98 artigos para o departamento de computação da UFMG e o algoritmo AREA 145 (dentre os agrupamentos do quadro de professores 3 tem elementos discutíveis) (seção 4.2),
- o algoritmo BIPARTITE encontrou 257 artigos para o quadro de professores da computação da Univ. Rice, o AREA 343 (seção 4.2).

Em resumo o resultado desta comparação para a área de Química (CHEM) foi:

- o algoritmo BIPARTITE encontrou 390 artigos para o quadro de professores da Unicamp e o AREA 1210 (sendo que apenas 1 dos agrupamentos de artigos de professores têm elementos discutíveis) 4.1,
- o algoritmo BIPARTITE encontrou 104 artigos para o quadro de professores da UFMG, o AREA 544 (onde 3 agrupamentos de artigos do quadro de professores têm elementos discutíveis) 4.1.

Tabela 4.1: Comparando os algoritmos de agrupamento BIPARTITE e AREA em relação ao número de artigos publicados no departamento de química na UFMG e na UNICAMP de 2003 a 2007

# de artigos		
Universidades	**BIPARTITE**	**AREA**
UFMG	104	644
Unicamp	390	1210

conjuntos de artigos por autor comparando AREA com BIPARTITE				
Universidades	**Maior(es)**	**Questionável(eis)**	**Igual(is)**	**Menor(es)**
UFMG	63	0	20	0
Unicamp	73	0	12	1

Tabela 4.2: Comparando os algoritmos de agrupamento BIPARTITE e AREA em relação ao número de artigos publicados no departamento de computação na Rice, UFMG e UNICAMP de 2003 a 2007

# de artigos		
Universidades	**BIPARTITE**	**AREA**
Rice	257	343
UFMG	98	145
UNICAMP	131	199

conjuntos de artigos por autor comparando AREA com BIPARTITE				
Universidades	**Maior(es)**	**Questionável(eis)**	**Igual(is)**	**Menor(es)**
Rice	19	0	17	0
UFMG	15	3	28	1
UNICAMP	21	4	26	1

4.3 Redução do volume de cálculo de similaridade usando a afiliação dos autores

As afiliações dos autores nos artigos também foram estudadas, mas esta foi uma atividade que custou um tempo razoável e pouco resultado diretamente prático para o problema. As afiliações, como está descrito no mecanismo de extração de dados na seção 3.2, não estão relacionadas uma a uma com os autores.

Como fato positivo é importante destacar que as afiliações são padronizadas em relação aos nomes dos Países. Outro fato positivo é que as afiliações dos Estados Unidos da América, que tem a maior produção de artigos atualmente, respeitam o formato nas frases de afiliação na parte de Sigla do Estado (NY, MA, etc.). Este fato é interessante, pois, no caso dos Estados Unidos, é possível trabalhar com os estados separadamente na hora de calcular os agrupamentos de afiliação. Um dos fatos negativos que se descobriu é que as frases de afiliações dos artigos brasileiros são as das mais confusas em relação à ordem dos campos, principalmente nos campos de estado e cidade, por aparecem invertidos ou misturados com outras informações. No entanto, ao utilizar uma tabela do IBGE das cidades brasileiras e o mecanismo de agrupamento usando similaridade de denominações foi possível identificar a cidade e seu estado correto na maioria dos casos. Outro fato extremamente negativo é que a maioria das informações de afiliação contida nos dados segue

uma estrutura até o nome da universidade ou instituição, mas muitas vezes a informação do departamento ou instituto se encontra misturada com outros campos.

A estrutura correta que é observada na maioria dos casos para frase de afiliação institucional de autor é: Grupo Institucional (Instituto, Departamento ou Laboratório), Instituição, Cidade, Estado e País.

Com estas análises foi possível identificar um problema no mecanismo de Consulta ao WOS, ou seja, o mecanismo para identificar as afiliações destacou algo que não havia sido percebido antes, ou seja, alguns artigos da USP não estão na base identificados como produção da USP. Estes artigos foram identificados como da USP não pelos mecanismos do extrator mas, indiretamente via o mecanismo de identificação de afiliação através de agrupamento por similaridade de nomes. No entanto, estes artigos estavam na base por terem como coautores membros de outras universidades que o extrator identificou corretamente. Quando se observa a *string* de consulta no anexo G é possível perceber que a consulta da USP não considera a palavra "USP", e acaba considerando vários artigos da UNESP (*Univ State Sao Paulo*) como produzidos pela USP. Isto não havia ficado claro inicialmente pois esta frase não aparecia na maioria dos artigos da USP.

Uma possível solução para o problema de identificar as afiliações dos artigos é consultar todos os artigos de cada país onde há universidades de interesse, e depois agrupar dentro do país por afiliação a cada instituição, (nos Estados Unidos, como a sigla do estado é respeitada o agrupamento pode ser realizado dentro de cada estado independentemente). No entanto, não houve tempo hábil para tratar deste aspecto.

Apesar deste problema os resultados finais não foram afetados drasticamente, visto que eles tiveram um bom desempenho em relação à precisão. A provável explicação para isto é que a consulta foi bastante permissiva na maioria dos casos e onde ela iria restringir erradamente acabou por não afetar os artigos de computação e química. No caso em que a consulta foi muito permissiva o mecanismo de redução de ambiguidade na identificação dos autores acabou por mitigar o problema priorizando a precisão dos resultados conforme pode ser visto na seção 4.6.

4.4 Levantamentos de valores limitantes superiores das estatísticas de artigos

Os valores limitantes superiores em relação às estatísticas de artigos são importantes para avaliar a qualidade dos resultados deste projeto, identificar com antecedência a importância relativa de instituições e áreas de pesquisa com base nesta visão abrangente da produção científica, e ainda permitir observar algumas características das principais entidades envolvidas.

Os cálculos destas informações estatísticas foram feitos utilizando-se os dados obtidos na WOS (seção 2.4) que foram carregados na base de dados descrita na seção 3.3. Estes dados se baseiam em publicações de artigos de 2003 a 2007.

Este levantamento foi agrupado em grupos de categorias de pesquisa de periódicos (seção 3.2), que são baseados nas definições de área de pesquisa afins (divididos em: CS, CHEM, PHYS e MATH) que foram identificados via consulta a especialistas de cada área e em outra definição destes grupos chamada de *Broad Subject Fields* (divisos em ENG, INTER, LIFE, MED, SCI e SOC) que é utilizada pelos *rankings* de universidades ARWU [12].

A quantidade de artigos e citações foi totalizada por universidade versus áreas de pesquisa e por universidade versus *Broad Subject Fields*. Tabelas separadas destacam a quantidade de artigos produzidos especificamente em periódicos, o que nos permite observar que nos dados utilizados há poucas ocorrências de artigos publicados que não sejam de periódicos em relação ao total de artigos.

Nos relatórios agrupados por áreas de pesquisa (CS, CHEM, PHYS e MATH), os dados foram ordenados de forma a destacar a importância das instituições brasileiras em produção científica (total de artigos e total de citações) especificamente na área de computação (CS). As instituições que não são brasileiras também estão ordenadas desta forma mas em separado e se encontram logo abaixo das universidades brasileiras. Para destacar estes critérios de classificação, foi colocada uma coluna identificando o país da instituição. Pelo mesmo motivo o título da coluna CS está em negrito e os dois primeiros colocados na ordenação decrescente em termos de produção científica em CS estão em negrito também. As instituições estrangeiras que tem produção inferior à primeira brasileira estão em itálico.

Analogamente nos relatórios agrupados por *Broad Subject Fields* (ENG, INTER, LIFE, MED, SCI e SOC), os dados foram ordenados de forma a destacar a importância das instituições brasileiras em produção científica (total de artigos e total de citações), especificamente em (ENG) Engineering/Technology and Computer Sciences.

Abaixo são listadas as tabelas como os resultados dos relatórios:

- Número de artigos por universidade e por área de pesquisa na Tabela D.1
- Número de artigos por universidade e por *Broad Subject Fields* na Tabela D.2
- Total de citações de artigos por universidade e por área de pesquisa na Tabela D.3
- Total de citações de artigos por universidade e por *Broad Subject Fields* na Tabela D.4
- Número de artigos em periódicos por universidade e por área de pesquisa na Tabela D.5

- Número de artigos em periódicos por universidade e por *Broad Subject Fields* na Tabela D.6
- Total de citações de artigos em periódicos por universidade e por área de pesquisa na Tabela D.7
- Total de citações de artigos por universidade e por *Broad Subject Fields* na Tabela D.8

Apenas observando estas tabelas é possível ver que a *USP* e a *Unicamp* são a primeira e a segunda colocada em produção científica dentre as instituições brasileiras tanto em área de pesquisa CS como em *Broad Subject Field ENG*.

Ao comparar a *USP* com as universidades estrangeiras na área de pesquisa *CS* é possível ver que ela supera em total de artigos as universidades de *NorthWestern*, *Duke*, *Yale*, *Rice* e *Syracuse*. No caso de total de citações a *USP* supera apenas as universidades de *Rice* e *Syracuse*.

Por outro lado, considerando o *Broad Subject Field ENG* na comparação entre a *USP* com as estrangeiras, se observa que esta supera em total de artigos as universidades *Caltech*, *Cornell*, *NorthWestern*, *Harvard*, *Princeton*, *Duke*, *Rice*, *Yale* e *Syracuse*. No caso de total de citações em *ENG* a *USP* supera apenas as universidades de *Yale* e *Syracuse*.

Em relação às estrangeiras as duas primeiras na área de pesquisa *CS* são as universidades *MIT* e *Berkeley*, tanto em total de artigos como em total de citações. Em *Broad Subject Field ENG* as duas primeiras são as universidades de *Texas Austin* e *Illinois*.

Analisando estes dados pode-se observar nos resultados de algumas universidades que na área de química (*CHEM*) a quantidade de artigos produzidos pela instituição como um todo em química se aproxima razoavelmente do total produzido pelo departamento ou instituto desta área de pesquisa da universidade, como no caso da UFMG e Unicamp nos dados obtidos na Tabela D.1 comparados com os resultados produzidos com o mecanismo de identificação de autores usando redução de ambiguidades (algoritmo *AREA*) na Tabela 4.2 para o quadro de professores destas universidades nestas áreas, tabulados nas Tabelas 4.1 e 4.1. Conforme estas tabelas o total de artigos produzidos nestas instituições na área de química pelo quadro professores do instituto ou departamento de química é: 644 (457 distintos) para UFMG e 1210 (1011 distintos) para Unicamp, que são valores respectivamente muito próximos aos 681 artigos para UFMG e 1825 artigos para Unicamp produzidos pelas instituições como um todo na área de química.

Usando este mesmo procedimento para área de computação nas mesmas instituições pode-se ver que a produção de artigos de Computer Science pelo quadro de professores na ciência da computação (*CS*) nestas três instituições é menos concentrada que na química (*CHEM*). Conforme se observa nos dados tabulados nos resultados do algoritmo *AREA*

nas Tabelas 4.2, 4.2, 4.2 e nos dados brutos totalizados por instituições na Tabela D.1 Com base nestes dados na Univ de Rice tem 619 artigos nas áreas próximas de ciência da computação (*CS*), enquanto apenas 343 (322 distintos) foram produção do quadro de professores da computação. No caso da *CS* para *Unicamp* 928 artigos, quando na produção de artigos do Instituto de Computação da *Unicamp* foram encontrados 199 (168 distintos), e na produção da *UFMG* foram 401 artigos comparado aos 145 (110 distintos) encontrados no seu departamento de *CS*.

As tabelas contendo os resultados do levantamento de valores limitantes superiores em relação às estatísticas de artigo encontram-se no apêndice D.

4.5 Rankings de universidades baseado em "prestígio"

Nesta seção são apresentados os resultados obtidos utilizando um dos produtos da análise da rede social de coautoria. Os resultados mostram que a medida estrutural de prestígio na rede social de coautoria produzida pela ferramenta foi capturada, ou seja, que existe significativa correlação entre o prestígio na rede de coautoria agregado pelo quadro de professores dos departamentos de computação para suas respectivas universidades em relação aos *rankings* de universidades por área de pesquisa computação e *broad subject field ENG* (Engenharia/Tecnologia e Ciência da Computação).

Um dos produtos da ferramenta é a rede de coautoria onde a metodologia utilizou as relações de coautoria e outros indícios de similaridade entre as denominações de autores e a lista do quadro de professores do departamento/instituto de computação de algumas das universidades elencadas nos principais *rankings* de universidades conhecidos. O objetivo de construir este *ranking* é mostrar que a metodologia de construção da rede de coautoria foi eficaz em produzir uma informação mais precisa do que a obtida sem considerar os indícios de similaridade indireta (emails, área de pesquisa e afiliações) e as relações de coautoria (que são informações sem relação estrita com as denominações de autores dos artigos) sendo capaz de capturar a informação de prestígio nesta rede social.

Na rede social de coautoria foi aplicado o algoritmo de *Pagerank* (na seção 2.3.2) que produz uma pontuação do prestígio dos atores desta rede, que no caso são o quadro de professores dos departamentos/institutos de computação das universidades (seção 3.5.1) e outros coautores que são de outras instituições. No caso de autores dos quais não são dadas informações de referência, como sua instituição, email e denominações foi possível considerar como indício de similaridade apenas as similaridades entre denominações abreviadas e de nome completo e indícios advindos puramente da relação de coautoria.

A maioria dos estudos de análise de redes sociais se baseia em dados nos quais existam relações estritas entre os atores da rede. No entanto, esta abordagem não é possível pela natureza dos dados que foram utilizados. Estes dados possuem um tanto de problemas de

ambiguidades (denominações de entidades) e algumas relações indiretas (lista de emails e lista de afiliações relacionadas diretamente ao artigo e não a cada autor). A metodologia obteve portanto uma amostra da rede social que respeita alguns princípios, ou seja, foram consideradas dentro da amostra apenas as relações entre autores que possuam fortes indícios de estar corretamente atribuída sua relação de autoria. Desta forma foi aplicada a análise social nesta amostra que resultou em um *ranking* do prestígio dos autores da rede de coautoria que pode ser usado para ordenar o prestígio das universidades que eles fazem parte e compará-los com alguns *rankings* das melhores universidades de computação ou engenharia que utilizam dados de fontes semelhantes ou critérios compatíveis aos deste experimento.

O cálculo do *Ranking* das universidades

O componente principal do grafo de coautoria abriga as relações de maior prestígio na rede de coautoria, desta forma foi considerado o resultado do algoritmo *PageRank*. O algoritmo do *PageRank* pontua cada autor a_i pertencente a este componente pelo principal pelo seu prestígio nesta rede R_{a_i} (seção 2.3.2). Se u é a universidade ao qual o autor i está afiliado então $a_i \in A_u$. No experimento foram considerados apenas os primeiros M pontuados, onde $M = 191$ particularmente. Para contabilizar a pontuação de cada universidade foi atribuído um peso a cada autor a_i que valoriza sua posição no *ranking* de prestígio na de rede de colaboração científica $M - R_{a_i}$, desta forma se a_1 possui posição $R_{a_1} = 1$, a ele é atribuído a pontuação $M - 1$, ou $191 - 1$. Este valor 190 irá se somar à pontuação da universidade u ao qual ele é afiliado. Portanto, a pontuação da universidade u é:

$$R_u = \sum_{a_i \in A_u} M - R_{a_i}, \tag{4.1}$$

onde:

A_u é subconjunto dos autores que são afiliados universidade u e
R_{a_i} é o ranking do prestígio do autor a_i na rede de coautoria

A Tabela 4.3 mostra as universidades ordenadas de forma decrescente em relação à sua pontuação e este *ranking* foi comparado aos da USNews 2010 para Computação, ARWU 2007 e 2008 para *broad subject field ENG* (Engenharia/Tecnologia e Ciência da Computação) e ARWU 2009 para Ciência da Computação. Para mais detalhes sobre estes

rankings veja a seção 2.8 e a Tabela 4.4. Os 191 autores mais pontuados em prestígio na rede de coautoria estão na Tabela F.1.

Tabela 4.3: Prestígio dos departamentos de computação baseado no prestígio na rede de coautoria do seu quadro de professores

Universidade	Brasileira	Pontuação	Posição
illinois	N	2955	1
mit	N	2341	2
cmellon	N	1437	3
waterloo	N	1381	4
berkeley	N	1360	5
stanford	N	1264	6
northwestern	N	1122	7
princeton	N	940	8
texas-austin	N	862	9
ufrj	S	678	10
yale	N	598	11
duke	N	552	12
rice	N	505	13
cornell	N	435	14
syracuse	N	366	15
caltech	N	331	16
unicamp	S	219	17
imperial	N	198	18
puc-rio	S	172	19
ufrgs	S	166	20
harvard	N	100	21
usp	S	18	22
ufmg	S	1	23

A Tabela 4.4 resume todos os *rankings* que foram comparados com o *ranking* dos departamentos e institutos de ciência da computação e engenharia da computação baseado no prestígio do seu quadro de pesquisadores na rede de coautoria.

Tabela 4.4: As posições de todos os *rankings* em relação ao do experimento (Prestígio na rede social de coautoria)

	Prestígio em coautoria	ARWU			USNews
	CS	ENG		*Comp.Sci*	*Comp.Sci*
Universidade	2003-2007	2007	2008	2009	2010
illinois	1	3	3	10	2
mit	2	1	1	2	1
cmellon	3	6	6	4	1
waterloo	4	13			
berkeley	5	4	4	3	1
stanford	6	2	2	1	1
northwestern	7	9	9		8
princeton	8	10	10	5	3
texas-austin	9	5	5	7	3
ufrj	10				
yale	11			12	6
duke	12	14		11	7
rice	13				6
cornell	14	7	8	6	2
syracuse	15				
caltech	16	8	7	8	4
unicamp	17				
imperial	18	11	11		
puc-rio	19				
ufrgs	20				
harvard	21	12	12	9	5
usp	22				
ufmg	23				

A correlação entre os *rankings* foi realizada utilizando o coeficiente de correlação de Kendall conforme descrito na seção 2.8.1. Um aspecto importante a ser lembrando é que a comparação é realizada em relação à ordem dos participantes por pontuação, portanto somente é necessário considerar os participantes que foram pontuados nos dois *rankings* sendo comparados. O coeficiente de correlação de Kendall τ_b desconta os empates ao calcular a proporção da concordância entre duas classificações e também no cálculo de sua variância.

Na Tabela 4.5 cada linha corresponde à correlação de um dado *ranking* com o que foi produzido pelo experimento, e nela são destacadas a identificação de cada um deles (autor e ano de publicação) e em seguida são apresentados os valores: a quantidade de universidades presentes em ambos os participantes da comparação (n), a dimensão da coincidência entre eles (S), o valor do denominador do coeficiente (D) que corresponde ao valor máximo possível para S, o valor do coeficiente de correlação kendall (τ_b), a variância de S ($\sigma_S{}^2$) e o nível de confiança de haver correlação entre os *rankings* (α).

Tabela 4.5: Resultados do coeficiente de correlação comparando outros *rankings* de universidades com o *ranking* produzido no experimento

Autor	Ano	Área ou *BSF*	n	D	S	τ_b	$\sigma_S{}^2$	α
AWRU	2007	ENG	14	91	41	0,4505	333,6667	0,0248
AWRU	2008	ENG	12	66	40	0,6161	212,6667	0,0061
AWRU	2009	Computer Science	12	66	24	0,3636	212,6667	0,0998
USNews	2010	Computer Science	14	86	34	0,3936	322,0000	0,0581

Nos resultados obtidos o nível de confiança(α) foi menor ou igual a $0,05\%$ em alguns dos casos (menor que $0,1\%$ em todos os casos) e houve correlação (τ_b de $0,3636$ a $0,6161$) com os *rankings* baseados em informações de prestígio do quadro de professores dos departamentos/institutos de computação obtidas em bases de dados (*ARWU*) e dos baseados em pesquisa de opinião (*USNews*) com correlação τ_b de $0,3936$. Como foi discutido na seção 2.8 os dois *rankings* utilizados se baseiam em dados relacionados direta ou indiretamente ao quadro de professores da área de pesquisa, seja através de pesquisa de opinião, ou seja, utilizando dados da ISI (seção 2.4). Os *rankings* que consolidam os resultados por *broad subject field ENG* tem por critério para selecionar os artigos que são contabilizados utilizando dados mais semelhantes ao utilizados no *ranking* do experimento. O *ranking* da *ARWU* para ciência da computação utiliza um critério diferente do utilizado pela ferramenta para atribuir um artigo a área de ciência da computação (veja seção 2.8) e também só iniciou sua publicação em 2009, o que pode explicar a correlação menor em relação à área temática ENG.

Outro aspecto importante a destacar é que a avaliação do experimento proposto neste trabalho considera os departamentos e institutos de ciência da computação engenharia da computação, e os outros *rankings* abrangem a produção científica nesta área dentro da universidade como um todo.

No levantamento feito na base de artigos produzidos por estas universidades de 2003 a 2007 foi encontrado um total de 35710 atribuídos à área de pesquisa ciência da computação (*CS*) e 51117 a *broad subject field ENG*.

Os resultados obtidos pela ferramenta identificaram como produção dos departamentos e institutos a quantidade de 5723 artigos para área de ciência da computação *CS* e 5440 artigos para *broad subject field ENG*. Esta proporção leva ao percentual de 10% da produção da universidade para *CS* e de 11% para *ENG*.

No aspecto do número de citações atribuídas para as universidades como um todo em pesquisa em ciência da computação (*CS*) resultou em 79840 e 154933 para *broad subject field ENG*. Os resultados obtidos pela ferramenta identificaram como produção dos departamentos e institutos 11230 citações para área de ciência da computação (*CS*) e 10934 citações para *broad research field ENG*. Assim, de forma semelhante, foi obtida a proporção de 14% do volume citações das universidades para *CS* em de 7% para *ENG*.

Outro aspecto é que da produção científica dos departamentos e institutos de computação uma parte significativa não é atribuída à área de computação ou a *broad subject field ENG*. O montante da quantidade de artigos foi de 6426 e 21873 citações, o que destaca o dobro de citações contribuindo para o desenvolvimento científico de outras áreas da ciência, mas o volume de artigos indica que 11% dos artigos produzidos não são da área de *CS* e contribuíram com 48% das citações e 16% dos artigos que não são de *broad subject field ENG* e contribuíram com 51% do volume de citações.

Na amostra das produções científicas relatadas nas páginas das universidades foram encontrados alguns cientistas que faziam parte do instituto de computação e também de institutos em outras áreas da ciência. Aparentemente isto contribuiu para o seu aumento de prestígio na rede de colaboração científica.

Esta abordagem de atribuir o prestígio dos pesquisadores na rede social de coautoria às instituições as quais eles são afiliados poderia ser utilizada para atribuir o prestígio destes cientistas aos veículos em que eles publicam seus artigos (periódicos e conferências). No entanto, não houve tempo hábil para executar este experimento.

4.6 A qualidade dos resultados

Uma amostra da produção científica do quadro de professores das 24 universidades que fizeram parte do experimento foi realizada manualmente visitando as páginas das universidades (as mesmas utilizadas para obter os nomes dos pesquisadores e seus emails que são dados de entrada da ferramenta). O procedimento adotado foi escolher aleatoriamente 384 professores dos 1512 que fazem parte do quadro de professores na área de computação e engenharia da computação de todas estas universidades. Apenas os artigos em periódicos foram coletados, visto a sabida deficiência da ISI no quesito artigos em conferências. A dificuldade de obter informações sobre produção científica na área de computação em relação a outras áreas é bem explorado em um artigo recente sobre avaliação das universidades brasileiras em qualidade educacional e produção científica [25].

Infelizmente não foi feito um levantamento apropriado de dados de teste. Este levantamento foi realizado parcialmente não sendo representativo para as universidades brasileiras. No entanto o volume foi razoável envolvendo dados de 174 pesquisadores em quantidade proporcional a lista de professores de Ciência da Computação ou Engenharia da Computação das universidades (berkeley, caltech, cmellon, cornell, duke, harvard, illinois, imperial, mit, puc-rio), resultando em um volume de 880 artigos atribuídos a estes pesquisadores.

Assim, foi obtida uma referência para realizar algumas medições comparativas de desempenho entre as fases da metodologia mas de pouco confiabilidade estatística para comparar com outros algoritmos semelhantes. Como foram obtidos dados de 174 pesquisadores dos 384 sorteados, faltaram 210 distribuídos proporcionalmente à lista de professores das universidades (northwestern,princeton, rice, stanford, syracuse, texas-austin, ufmg, ufpe, ufrgs, ufrj, unicamp, usp, waterloo, yale). No entanto, estima-se que um total de artigos em quantidade menor falte para terminar o levantamento pois há muitas universidades brasileiras nesta lista.

As medidas de desempenho micro-agregado para recuperação, precisão e consequentemente a *medida-F* descritas na seção 2.7.5 serviram para comparar as fases dos algoritmos.

A medida F_1 é usualmente utilizada para medir o desempenho de algoritmos de categorização de textos e desta foram obtidos valores comparáveis aos destes sistemas, no valor de 0,8306 (precisão 0,9889 e recuperação 0,7159) obtidos na fase do aumento da precisão via *rede de coautoria inicial*, após a *fase 4* da metodologia – *rede de coautoria extensa* o desempenho F_1 atingiu o valor de 0,8360 e usando $F_{0,5}$ que prioriza o dobro da precisão em relação à recuperação o resultado foi de 0,9188 contra 0,8906 se não utilizada a rede de coautoria na metodologia – *fase 2*. Estes dados de desempenho dos algoritmos foram calculados utilizando os conceitos apresentados na seção 2.7.5 e estão resumidos nas tabelas abaixo na forma das tabelas de contingência consolidadas (Tabela 4.6) para cada fase do algoritmo descrito na seção 4.6 e os resultados das medidas de desempenho médio micro-agregado na Tabela 4.7.

Tabela 4.6: Tabelas de contingência consolidadas (todas as categorias da amostra) de cada fase da metodologia e do algoritmo bipartite

Algoritmo	*a*	*b*	*c*	*d*
FASE 4	632	7	241	9401
FASE 3	625	7	248	9408
FASE 2	879	135	0	9026
BIPARTITE	E831	136	48	9073

Tabela 4.7: A melhoria do desempenho priorizando a precisão em dobro (medida $F_{0,5}$ - Micro agregado)

Algoritmo	R	P	F	A	E	medida F_1	**medida** $F_{0,5}$
FASE 4	0,7239	0,9890	0,0007	0,9993	0,0247	0,8360	0,9216
FASE 3	0,7159	0,9889	0,0007	0,9993	0,0254	0,8306	0,9188
FASE 2	1,0000	0,8669	0,0147	0,9866	0,0134	0,9287	0,8906
BIPARTITE	0,9454	0,8594	0,0148	0,9865	0,0183	0,9003	0,8753

O erro na identificação da autoria foi de 2%, o que levanta questionamentos sobre o uso desta metodologia para comparar quantidades de publicações ou de citações entre pesquisadores, ou grupos de pesquisadores, pois na área de computação foi encontrada uma média de 5 publicações em periódicos por autor. Na comparação da produção de departamentos ou institutos de computação e engenharia da computação de universidades brasileiras também não seria recomendável, pois a maior produção de artigos de 2003 a 2007, encontrada nos resultados, foi de 202 artigos para a que mais produziu. No caso das universidades estrangeiras a maior produção foi de 812 artigos. Outro aspecto a considerar é a grande produção que é realizada em conferências que ainda não estão inseridas nas bases da WOS, e de outras bases semelhantes.

Uma dificuldade que se apresentou foi a comparação entre os resultados de medidas de quantidades de publicação e de citações com os *rankings* divulgados atualmente, pois muitos destes consideram os artigos produzidos por cada universidade como um todo na maior parte deles. Alguns trazem resultados separados por área e por área temática de pesquisa mas nem todos divulgam os detalhes de como é definido um campo de pesquisa ou uma área de pesquisa. Sendo assim, estes fatos estabelecem incompatibilidades com os experimentos realizados com a ferramenta, que tiveram como objetivo o pequeno grupo de pesquisadores dos institutos ou departamentos de computação e engenharia da computação destas instituições, e como pode ser visto no levantamento dos dados da base na seção E as publicações destes grupos em relação ao todo da instituição correspondem a uma participação bem menor do que o montante total. Ou seja, de toda a produção científica dos departamentos e institutos de computação uma parte significativa não é atribuída à área de computação ou a *broad subject field ENG*. O montante da quantidade de artigos foi de 6426 e 21873 citações, o que destaca o dobro de citações contribuindo para o desenvolvimento científico de outras áreas da ciência, mas o volume de artigos indica que 11% dos artigos produzidos não são da área de *CS* e contribuíram com 48% das citações e 16% dos artigos que não são de *broad subject field ENG* e contribuíram com 51% do volume de citações.

Portanto, isso compromete a comparação com os *rankings* SCIMago [19] e da Times [18], apesar de eles dividirem os *rankings* por campo de pesquisa e área de pesquisa, mas não divulgam detalhes de como estes são definidos, e consideram a produção científica da instituição como um todo.

Mas para os *rankings* que utilizam critérios compatíveis com o *ranking* proposto neste trabalho, foi encontrada uma semelhança significativa entre eles, a menos das universidades brasileiras que poucos destes *rankings* as levam em consideração. O coeficiente de correlação de *rankings Kendall* obteve um valor significativo comparando aos *rankings* que focam em áreas de pesquisa. Utilizando a interpretação de probabilidade de correlação entre *rankings* proposta no artigo [23] que considera a proporção da correlação positiva obtida, lembrando que a correlação varia de -1 ordem inversa a 1 concordância total:

$$\hat{p} \;=\; \frac{\tau + 1}{2}$$

Portanto, utilizando esta medida para interpretar a proporção de correlação entre o *ranking* produzido pela ferramenta com o *ARWU* foi obtido de 68% para *ENG* 2008 e de 80% para *CS* 2009. Ao comparar com o *ranking USNews* para *CS* em 2010 o resultado foi de 69%, o nível de confiança destes resultados está na Tabela 4.5. Isto demonstra que a proposta de um *ranking* baseado em status na rede de coautoria captura uma medida da estrutura das relações de prestígio que se mantém no decorrer dos anos, apesar das variações que ocorreram do decorrer dos anos nestes *rankings*.

Observando os componentes da rede de coautoria foi possível identificar a diferença na atividade de colaboração entre pesquisadores de áreas de estudo aplicadas e as teóricas. Os autores na área de teoria ficaram em posição inferior no prestígio e alguns ficaram fora do componente principal da rede, por vezes, o fato de produzirem poucos artigos ou com poucos coautores adicionados a baixa similaridade de denominações levou a agravar este aspecto.

A possibilidade de avaliar o prestígio dos veículos de divulgação científica e instituições de pesquisa tem se destacado como de interesse de toda a comunidade científica e dos tomadores de decisão em relação às estratégias de investimento, a exemplo da popularidade do *JCR*. No entanto, a melhor forma de medir esta informação, como outras de utilidade para apoiar o desenvolvimento da ciência, depende de maior transparência da forma de registrar a produção científica conforme opina Lane em seu recente artigo na Nature [26]. Outro aspecto a considerar é a grande produção que é realizada em conferências que ainda não estão inseridas nas bases da WOS, e de outras bases conforme é discutido por Przibisczki na sua tese [15]. Algum progresso já foi feito em relação a identificar unicamente os artigos publicados e exemplo da ISI que, com isto, consegue produzir o JCR (Jour-

nal Citation Ranking). No entanto, como foi visto nos desafios enfrentados do presente trabalho é necessário identificar unicamente os pesquisadores, como a exemplo da Plataforma Lattes, e que também identifica os grupos e projetos de pesquisa nelas cadastrados. No texto a seguir são apresentados recentes progressos neste desafio, valorizando alguns aspectos da Plataforma Lattes.

Plataforma Lattes

A Plataforma Lattes é uma arquitetura de informações em CT&I desenvolvida para o CNPq gerir suas atividades de fomento e para integrar em um mesmo ambiente os diversos atores ligados ao sistema nacional de inovação no país. Além de viabilizar a interoperacionalidade dos sistemas de informação das agências federais, a Plataforma Lattes tem racionalizado o processo de gestão de CT&I [11]. A estrutura arquitetônica da Plataforma Lattes é composta de níveis conceituais (camadas) levados à prática por meio de instrumentos e métodos, que compreendem desde o arquivo de dados sistematizados nas unidades de análise até a extração de conhecimento referente à informação nacional, sobre CT&I. As informações de currículos, grupos de pesquisa e projetos em CT&I realizadas no país são públicas e permitem a extração de novos conhecimentos sobre C&T. Para permitir estes tipos de análises alguns aspectos são importantes na Plataforma Lattes:

DOI Um identificador único de documento (*Document Object Id*). Nem todas as *DLs* garantem unicidade de documento. Mesmo a Plataforma Lattes está tratando este aspecto agora. Outros sistemas criaram seus próprios mecanismos, mas os princípios do protocolo; *DOI* são usados com referências nestas várias iniciativas da comunidade *open-source* e da comunidade de publicações. Este protocolo se tornou o padrão de fato para identificação de documentos. A Plataforma Lattes recentemente vem atualizando as publicações dos pesquisadores com os respectivos *DOI*;

RCID Um identificador único de pesquisador e contribuinte (*Research and Contributor ID*) Algumas iniciativas recentes de âmbito nacional nos Estados Unidos foram [26]: *ORCID - Open Research and Contributor ID* e um projeto lançado em Dezembro de 2009 envolvendo *Thomson Reuters* e *Nature Pub. Group.*;

MODELO Um padrão para relatar progressos científicos. A Plataforma Lattes tem sido exemplo neste aspecto. O *US National Science and Tecnology Concil* recomendou um padrão para relatar o progresso científico;

RGID Um identificador de Grupo de Pesquisa. O Lattes tem esta iniciativa, mas existem alguns artigos que tem avaliado características da atividade de pesquisa destes grupos cadastrados no Lattes. Na plataforma Lattes, juntamente com o identificador único de pesquisador/colaborador, tem permitido gerar resultados sem riscos de

perda de qualidades que estas ambiguidades podem causar. No entanto, ainda falta ao Lattes garantir a identificação única dos documentos e dos coautores;

RPID identificador de Projeto de Pesquisa O Lattes tem esta iniciativa;

COLETA Acesso aberto aos dados. No caso do Lattes este acesso é restrito para consultas públicas.

Comentários finais

Mesmo o bom exemplo da Plataforma Lattes não progrediu no sentido de identificar nos artigos de cada pesquisador os coautores que estão cadastrados na própria base de dados. Outro aspecto importante é oferecer estes dados de forma estruturada para acesso, para que estudos de novas métricas de análise de redes sociais possam ser realizados. Com toda certeza algum compromisso ético deve ser firmado com quem for utilizar estes dados. O artigo Mena-Chalco e Cesar-Jr [33] também confirma o valor das ideias incorporadas na Plataforma Lattes, assim como o desafio que tiveram para implementar uma ferramenta de extração e análise dos dados bibliográficos, disponíveis nesta base de dados.

Outro aspecto foi o grande desafio apresentado no projeto em relação aos dados obtidos na consulta da base bibliográfica em relação à situação na qual houvesse acesso aos dados conforme são estruturados internamente na ISI (seção 2.4). Na grande maioria dos artigos desta base a afiliação dos autores às universidades está estritamente definida, ou seja, existe uma relação estrita em o (nome abreviado, nome completo se houver) e a instituição a qual este autor credita a sua afiliação, e quando ao seu email está presente e existe também a relação estrita com o seu email. Assim, a ferramenta tendo acesso a estes detalhes, agregaria um volume maior de evidências da autoria dos artigos e então a atividade de desambiguação se restringiria aos nomes dos autores e não haveria incertezas quanto ao proprietário do email e a afiliação a instituição, sendo equivalente ao problema tratado no artigo 2.5.

Além disso, possuir um conjunto de dados completo com todas estas informações definidas de forma estrita seria o ideal para realizar experimentos em algoritmos de agrupamento, categorização vértices de grafos parcialmente categorizados e técnicas de aprendizagem supervisionada, semi supervisionada ou não supervisionada. No entanto, a obtenção manual da relação de autoria nestes dados é um processo trabalhoso.

A ferramenta apresentada neste trabalho sugere com os resultados obtidos a possibilidade de servir a esta atividade de categorizar uma base de dados de dados bibliográficos de tamanho razoável, ou seja, utilizando-a como ferramenta de auxílio à tomada de decisão ao categorizar cada autoria dos artigos contribuindo para acelerar o processo de ter estes dados estritamente definidos.

Capítulo 5

Conclusões e trabalhos Futuros

O presente trabalho aborda problemas que não foram totalmente resolvidos em sistemas importantes de acesso à base de dados bibliográficas, como WOS e Scopus, com relação à identificação única dos autores dos artigos. Para tratar este problema, foi utilizada a análise da estrutura das redes sociais para solução/redução de ambiguidades em identidades de uma base de dados bibliográficos, e desta forma produzir um mecanismo mais preciso de sumarização de informações nos dados bibliográficos. Sendo assim, de forma prática, foi implementada uma ferramenta que permite algumas comparações da produção científica entre instituições, departamentos, áreas de pesquisa e entre países. Além disso, o mecanismo de agrupamento e as funções de similaridade de textos foram utilizados por outro trabalho de Mestrado[15] para agrupar nomes de veículos de publicações científicas.

Os experimentos utilizaram um grande volume de dados que exigiu um bom espaço de armazenamento de disco, de memória RAM e de processamento. Deve-se considerar ainda que foi necessário executar variações da metodologia, para que fosse possível comparar o desempenho entre elas, o que acabou demandando ainda mais recursos de máquina.

Nas seções anteriores foram discutidas metodologias que permitiram a viabilidade da execução dos experimentos frente ao volume de dados. No entanto, o desafio em conhecer a estrutura dos dados da WOS, mas também, a aplicação de técnicas apropriadas para que o volume de dados fosse tratado pelas fases mais custosas do processo demandou mais tempo do que havia sido planejado. Além disso, se despendeu um considerável tempo para levantar um pequeno conjunto de dados, bem como categorizá-los para ser possível a comparação do desempenho entre as metodologias.

Através dos experimentos, foi possível concluir que apesar de um mecanismo de agrupamento simples, e de funções de similaridade com definições de parâmetros apenas intuitivos, foi possível um bom resultado ao recuperar as informações exatamente por capturarem dados pertinentes. Foi possível mostrar também que, a estrutura da rede de coautoria pode ajudar a melhorar os resultados principalmente no aspecto da precisão em

detrimento da medida da recuperação. Os resultados, quando comparados com a amostra manualmente categorizada, mostram que os indícios de similaridade de baixa qualidade, nos dados que foram descartados e os que foram considerados, foram reforçados pela estrutura da rede de coautoria ou tinham boa qualidade de indícios de similaridade. Os indícios de similaridade tiveram um papel essencial no sucesso desta abordagem, relembrando que, é como foram chamadas as informações sem relação unívoca direta com as denominações dos autores, mas apenas ao artigo (lista de emails de autores do artigo, lista de afiliações dos autores do artigo, lista de categorias de pesquisa do artigo, nome do veículo).

O fato de a função de similaridade utilizada para o agrupamento de autores basear-se principalmente no nome do autor e no seu nome abreviado, se deve aos outros atributos não terem uma relação direta entre eles e os autores. Como a lista de denominações (a maioria com os emails) dos professores dos quadros das universidades é dado de entrada, isto permitiu que a metodologia conseguisse utilizar estes indícios de similaridade no processo. Após as denominações dos autores já estarem agrupadas apenas por similaridade de nomes, ou seja, para as denominações de autores semelhantes à de um pesquisador da instituição a que faz parte e que está na lista de afiliações do artigo, foi priorizado atribuí-lo ao autor desta instituição, descartando as denominações semelhantes a este pesquisador, que não possuíam na lista de afiliações de autores do artigo a instituição que este pesquisador era afiliado. Como também, o caso de neste conjunto de denominações existir uma denominação que está relacionada a um artigo, cuja lista de emails tenha um email deste pesquisador, reforça a similaridade com este autor.

Outro indício de similaridade entre autores que teve importante participação nos resultados, é baseado na busca de um caminho de coautoria que, partindo de uma denominação com similaridade de nomes apenas razoável e com poucos indícios de ser daquele autor, e passando pelas relações de coautoria, retorna a outra denominação com muitos indícios de ser uma atribuição correta a este autor. Este mecanismo ajuda a selecionar os dados de maior confiabilidade aumentando a precisão dos dados. No aspecto do uso destes atributos que não implicam percurso na rede social na função de similaridade, poderia abrir discussões sobre o uso de outras técnicas, como otimização combinatória, mas neste trabalho não se aprofundou neste outro enfoque, mas seria assunto de interesse para trabalhos futuros.

A comparação dos resultados de medidas de quantidades de publicação e de citações com os *rankings* divulgados atualmente se mostrou uma tarefa difícil, pois muitos deles consideram os artigos produzidos por cada universidade como um todo. Alguns trazem resultados que são separados por área e por área temática de pesquisa, mas não há um consenso na definição de áreas de pesquisa. Sendo assim, devido ao foco de os experimentos ser o pequeno grupo de pesquisadores dos institutos ou departamentos de computação

destas instituições, e como foi verificado no levantamento dos dados da base que as publicações destes grupos em relação ao todo da instituição, observou-se uma participação bem menor do que o montante total resultando numa incompatibilidade com os *rankings* que consideram dados da universidade como um todo.

Uma estratégia que não foi utilizada no trabalho, é a de dividir o processamento do agrupamento por universidades (da lista de universidades cujos nomes do quadro de professores que é dado de entrada para a ferramenta). Esta abordagem foi implementada inicialmente, mas como pode ser visto nos exemplos no anexo H.4 ocorre perda de informações importantes, como o caso de um autor que não pertence a nenhuma destas universidades das quais se possa identificar a afiliação do artigo (que não conste na lista dada como entrada para ferramenta). Este caso ficaria fora do agrupamento, e como consequência, no momento de criar a rede de coautoria, não seria possível encontrar alguma relação de coautoria das quais este autor participa. Nos exemplos são mostrados alguns casos em que estas relações coautoria foram importantes para reforçar os indícios de similaridade entre denominações de autores das universidades que foram escolhidas para os experimentos.

Observando os componentes da rede de coautoria, foi possível identificar a diferença na atividade de colaboração entre pesquisadores de áreas de estudo aplicadas e as teóricas. Os autores na área de teoria ficaram em posição inferior no prestígio e alguns ficaram fora do componente principal da rede. Por vezes, o fato de produzirem poucos artigos ou com poucos coautores adicionados a baixa similaridade de denominações levou a agravar este aspecto.

Para demonstrar estes conceitos foi utilizada como fonte de dados bibliográficos um banco de dados extraído da WOS ISI [39] das universidades de maior renome dentre as brasileiras e estrangeiras. Os dados se referem aos artigos produzidos de 2003 a 2007, e foram extraídos no início de 2008. O grupo de autores escolhido foi o quadro de professores dos departamentos de computação e engenharia da computação destas universidades, com vistas à comparação quantitativa e qualitativa entre estes grupos.

Uma amostra foi categorizada manualmente sorteando da lista de 1512 pesquisadores dessas universidades dos departamentos de computação e engenharia da computação um total de 384 deles. Como foi discutido na parte sobre a qualidade dos resultados na seção 4.6 não foi feito um levantamento apropriado de dados de teste, este levantamento foi feito parcialmente não sendo representativo para as universidades brasileiras. No entanto, o volume foi razoável, envolvendo dados de 174 pesquisadores em quantidade proporcional a lista de professores de Ciência da Computação ou Engenharia da Computação das universidades (berkeley, caltech, cmellon, cornell, duke, harvard, illinois, imperial, mit, puc-rio), resultando em um volume de 880 artigos atribuídos a estes pesquisadores. Mas esta amostra foi utilizada para realizar algumas medições comparativas de desempenho

entre as fases da metodologia.

As medidas de desempenho micro-agregado para recuperação, precisão e a medida-F foram utilizadas para comparar as fases dos algoritmos. A medida F_1 usualmente utilizada para medir o desempenho de algoritmos de categorização de textos obteve valores comparáveis aos destes sistemas, com um valor de 0,8306 (precisão 0,9889 e recuperação 0,7159) obtidos na *fase 3* indicando um aumento da precisão via rede de coautoria. Após a *fase 4* da metodologia atingiu o valor de 0,8360 para medida F_1 e usando $F_{0,5}$ que prioriza ao dobro a precisão em relação à recuperação o resultado foi de 0,9216 contra 0,8906 se não fosse utilizada a rede de coautoria.

O erro na identificação da autoria foi de 2%, o que levanta questionamentos sobre o uso desta metodologia para comparar quantidades de publicações ou de citações entre pesquisadores, ou grupos de pesquisadores, pois na área de computação foi encontrada uma média de 5 publicações em periódicos por autor. Na comparação da produção de departamentos ou institutos de computação de universidades brasileiras, também não seria recomendável, pois a maior produção de artigos de 2003 a 2007 encontradas nos experimentos foram de 202 artigos para aquela que mais produziu. No caso das universidades estrangeiras a maior produção foi de 812 artigos. Outro aspecto a considerar, é a grande produção que é realizada em conferências que ainda não estão inseridas nas bases da WOS, e de outras bases.

A rede de coautoria produzida pela ferramenta foi utilizada para produzir um *ranking* de universidades baseado no prestígio dos pesquisadores destas universidades nesta rede social. O coeficiente de correlação de *rankings Kendall* obteve um valor significativo quando comparando aos *rankings* que focam em áreas de pesquisa, como o ARWU [12] para campo área de pesquisa engenharia/tecnologia e ciência da computação dos anos 2007 e 2008 e para área de ciência da computação de 2009, e com o *ranking* USNews [31] para área de ciência da computação de 2010, que medem a reputação dos departamentos e institutos da computação e engenharia. Isto demonstra que o *ranking* baseado em status na rede de coautoria captura uma medida estrutural das relações de prestígio e que ainda se mantém no decorrer dos anos, apesar das variações que ocorreram de ano para ano em cada *ranking*. Ou seja, a rede de coautoria foi capaz de capturar uma informação de prestígio semelhante a estes *rankings*.

Uma sugestão para outros experimentos seria percorrer as redes de autoria em relação aos veículos de publicações, algo que poderia utilizar o conceito de "colégio virtual", encontrado nas redes de colaboração científica, em que pesquisadores de mesmos interesses de pesquisa colaboram inter-institucionalmente, e que geralmente publicam em alguns veículos de maior evidência com maior frequência.

Outros aspectos interessantes do modelo de dados e das ferramentas reunidas no sistema é que permitem visualizar diversas outros resultados além das medidas em contagem

de artigos dos autores, mas avaliar a estrutura de redes sociais como a de coautoria, e redes de *one-mode* produzidas através das redes de afiliação. Neste trabalho foi analisada a estrutura de uma rede de coautoria e foi utilizado um algoritmo de *ranking* de influência para destacar as instituições mais influentes. Mas Wasserman [45] sugere uma abordagem interessante para análise de redes de afiliação transformando-as em redes *one-mode*, que no caso poderia permitir colocar autores, relações de coautoria, e também, a relação dos autores com as instituições as quais ele pertence numa mesma rede, e desta forma analisar a influencia tanto dos autores, como das instituições num mesmo *ranking*. Outra possibilidade, seria fazer algo análogo ao que foi feito para atribuir influência dos pesquisadores na rede de coautoria para as suas instituições, trocando o objetivo da análise no sentido de atribuir importância aos veículos de publicações relacionados aos autores que nelas publicam, baseada na influência destes na rede de coautoria.

A comparação dos resultados de medidas de quantidades de publicação e de citações com os *rankings* divulgados atualmente se mostrou uma tarefa difícil, pois muitos deles consideram os artigos produzidos por cada universidade como um todo. Alguns trazem resultados separados por área e por área temática de pesquisa. Mas a dificuldade dessa comparação se agrava, ainda mais, devido ao foco dos experimentos ser o pequeno grupo de pesquisadores dos institutos ou departamentos de computação destas instituições, e como foi verificado no levantamento dos dados da base que as publicações destes grupos em relação ao todo da instituição, há uma participação bem menor do que o montante total.

Após a construção da rede social de autoria foi obtida uma informação interessante com base nos resultados para cada grupo de pesquisadores por universidade, onde uma quantidade significativa destes artigos produzidos por estes pesquisadores não serem de computação, mostrando o quanto eles têm contribuído para outras áreas de conhecimento com um percentual de 44 a 98% das citações sendo em artigos destes pesquisadores no campo da ENG (*Engineering/Technology*) e de 76 a 96% da quantidade de artigos. Na área de computação, também foi encontrado um percentual de 48 a 100% das citações para área e de 81 a 99 % da quantidade de artigos. Mas os *rankings* se assemelharam aos primeiros colocados e aos últimos destes *rankings*, a menos das universidades brasileiras que em poucos destes *rankings* as levam em consideração.

Inicialmente, também foi feito uma parte do estudo para química utilizando o mecanismo de agrupamento, mas este ficou incompleto, e não foi feito um levantamento de um conjunto de testes. No entanto, a expectativa é que a mesma metodologia como um todo possa ser aplicada para outras áreas de pesquisa. Sendo esta uma das oportunidades para trabalhos futuros.

A metodologia proposta poderia evoluir para um mecanismo que ajudasse a identificação de autoria de artigos para qualquer área de pesquisa usando bases de dados

bibliográficos, utilizando as informações da estrutura destas redes sociais. As análises de redes sociais poderiam ter maior confiabilidade, se houvesse uma amostra de dados para verificação, ou que houvessem iniciativas discutidas no trabalho para que as entidades (artigos, autores, pesquisadores, grupos de pesquisadores, instituições) fossem unicamente identificadas, e que mecanismos abertos de acesso aos dados fossem disponíveis usando os já estabelecidos protocolos de coleta. O exemplo da Plataforma Lattes já é reconhecido internacionalmente, mas é preciso avançar mais no sentido de garantir informações que possam ser relacionadas (coautoria por exemplo) e de livre acesso, como foi discutido na seção 4.6.

Apêndice A

Campos do registro de informações bibliográficas de artigos da WOS

Em cada registro da base de informações bibliográficas de artigos da WOS 2.4 há 38 colunas nomeadas por uma sigla de 2 letras. Na tabela abaixo são listados os significados de algumas destas colunas.

Tabela A.1: Campos do registro de informações bibliográficas de artigos

Col	Sigla	Descrição	Domínio
0	PT	Tipo de Publicação	J = Periódico; S = Evento
1	AU	Autores	Separados os autores por ";" no formato Sobrenome"," Primeira Letra de Nomes abreviados com "." e separados por "branco"
2	AF	Autores com primeiro nome não abreviado	Separados os autores por ";" no formato Sobrenome"," Primeiro Nome "branco" e Primeira letra de Nomes abreviados com "." e separados por "branco"
3	CA		
4	TI	Título	Texto
5	SO	Periódico/Evento	Periódico se PT=J, Evento se PT=S

Tabela A.1: continuação

Col	Sigla	Descrição	Domínio
6	SE	Veículo	Somente presente se PT $\neq$ J
7	LA	Língua	
8	DT	Tipo de Documento	Article, Bibliografical-ilem, Book Review, Correction, Editorial Material, Letter, Meting Abstract, Review
9	DE	Palavras Chaves	palavras separadas por ";"
10	ID	Palavras Chaves (categorias)	palavras separadas por ";"
11	AB	Resumo	Texto
12	C1	Afiliações	Afiliações separadas por ";" mas não ordenadas
13	RP		Nome e afiliação de um dos autores, o principal
14	EM	Emails	emails dos autores separados por ";" mas não ordenados
15	CR		
16	NR	Número de referências neste artigo	Números de itens nas referências deste artigo
17	TC	Número de citações a este artigo	número
18	PU	Editoras	
19	PI	País da Editora	
20	PA	Endereço da editora	
21	SN	CEP da editora	
22	J9	Periódico	

Tabela A.1: continuação

Col	Sigla	Descrição	Domínio
23	JI	Periódico	
24	PD	Data ou Mês da publicação	
25	PY	Ano da Publicação	ano em 4 dígitos
26	VL	Volume	número
27	IS		data
28	PN	Part Number	
29	SU	Suplemento	
30	SI		
31	BP	Página	
32	EP	Página	
33	AR	Código	
34	DI		
35	PG		
36	SC	Área de Pesquisa	frases separadas por ";" representando as Áreas de Pesquisa. Vide B.1 para saber os valores possíveis
37	GA		código
38	UT	Identificador único de publicação	identificação única para cada artigo na base da WOS

Apêndice B

Lista de categorias de área de pesquisa de periódicos

A tabela abaixo mostra as possíveis frases que representam categorias de áreas de pesquisa utilizadas no item (SC) (lista de categorias de área de pesquisa).

Tabela B.1: Categorias de área de pesquisa

ACOUSTICS
AGRICULTURAL ECONOMICS & POLICY
AGRICULTURAL ENGINEERING
⋮
TROPICAL MEDICINE
UROLOGY & NEPHROLOGY
VETERINARY SCIENCES
VIROLOGY
WATER RESOURCES
ZOOLOGY

Apêndice C

Categorias de áreas de pesquisa e suas áreas afins

Os artigos são classificados em áreas afins para áreas de pesquisa: *computer science*, *chemistry*, *physics* e *mathematics*, conforme as tabelas a seguir que as relacionam com categorias de pesquisa de periódicos. Esta informação foi levantada junto a especialistas destas áreas.

Tabela C.1: Categorias de área de pesquisa em periódicos para *Computer Science* (CS) e áreas afins

Computer Science – CS
Categoria de Pesquisa de Periódico
Computer Science
Computational
Informatics
Automation & Control Systems
Engineering, Electrical & Electronic
Mathematical
Mathematics
Operations Research
Robotics
Statistics
Probability
Telecommunications

Tabela C.2: Categorias de área de pesquisa em periódicos para *Mathematics* (MATH) e áreas afins

Mathematics – MATH
Categoria de Pesquisa de Periódicos
Mathematics
Mathematical
Operations Research
Computational
Computer Science
Statistics
Probability

Tabela C.3: Categorias de área de pesquisa em periódicos para Physics (PHYS) e áreas afins

Physics – PHYS
Categoria de Pesquisa de Periódicos
Physics
Chemistry, Physical
Materials Science, Multidisciplinary
Instruments & Instrumentation
Nuclear Science & Technology
Engineering, Electrical & Electronic
Mathematics
Optics
Astronomy
Astrophysics

Tabela C.4: Categorias de área de pesquisa em periódicos para *Chemestry* (CHEM) e áreas afins

Chemestry – CHEM
Categoria de Pesquisa de Periódicos
Chemistry
Materials Science
Instruments & Instrumentation

Apêndice D

Valores limitantes superiores em relação às estatísticas de artigos

Neste apêndice são apresentados os limitantes superiores em relação às estatísticas de artigos na forma de tabelas. Estas tabelas foram agrupadas por grupos de categorias de pesquisa de periódicos 3.2, que são baseados nas definições de área de pesquisa afins (divididos em: CS, CHEM, PHYS e MATH) que foram identificados via consulta a especialistas de cada área. As tabelas também foram agrupadas em outra definição destes grupos chamada de *Broad Subject Fields* ou áreas temáticas (divisos em ENG, INTER, LIFE, MED, SCI e SOC) que é utilizada pelo ranking de universidades ARWU [12].

Totalizou-se a quantidade de artigos e citações por universidade versus áreas de pesquisa e por universidade versus *Broad Subject Fields*. Tabelas separadas destacam a quantidade de artigos produzidos especificamente em periódicos, o que nos permite observar que nos dados utilizados há poucas ocorrências de artigos publicados que não sejam de periódicos em relação ao total de artigos.

Nos relatórios agrupados por áreas de pesquisa (CS, CHEM, PHYS e MATH), os dados foram ordenados de forma a destacar a importância das instituições brasileiras em produção científica (total de artigos e total de citações) especificamente na área de computação (CS). As instituições que não são brasileiras também estão ordenadas desta forma, mas em separado e se encontram logo abaixo das universidades brasileiras. Para destacar estes critérios de classificação, foi colocada uma coluna identificando o país da instituição. Pelo mesmo motivo o título da coluna CS está em negrito e os dois primeiros colocados na ordenação decrescente em termos de produção científica em CS estão em negrito também. As instituições estrangeiras que tem produção inferior à primeira brasileira estão em itálico.

Tabela D.1: Número de artigos por universidade e por área de pesquisa

Universidade	Brasileira	País	Todas	CHEM	**CS**	MATH	PHYS
usp	**S**	**BRASIL**	16188	2792	**1393**	1194	4025
unicamp	**S**	**BRASIL**	7261	1825	**928**	700	2297
ufrj	S	BRASIL	6200	1070	837	717	1702
ufmg	S	BRASIL	3747	681	401	308	893
ufrgs	S	BRASIL	4167	618	390	336	980
ufpe	S	BRASIL	1721	384	361	315	597
puc-rio	S	BRASIL	991	225	326	255	481
unb	S	BRASIL	1743	301	265	223	627
texas-austin	**N**	**EUA**	42143	2768	**4183**	3045	6471
illinois	**N**	**EUA**	23584	2783	**3865**	2929	6409
mit	N	EUA	13852	2043	2969	2155	6390
berkeley	N	EUA	20097	2649	2901	2235	7593
stanford	N	EUA	19782	1381	2533	1825	4379
waterloo	N	CANADA	6332	793	2203	1655	2373
cmellon	N	EUA	5575	752	2091	1749	2242
imperial	N	INGLATERRA	16353	2124	1872	1462	4603
harvard	N	EUA	39728	1438	1768	1553	3168
princeton	N	EUA	8917	847	1597	1356	4677
cornell	N	EUA	17253	1287	1548	1199	3251
caltech	N	EUA	10565	1236	1516	948	6290
northwestern	*N*	*EUA*	13092	1723	*1108*	931	3092
duke	*N*	*EUA*	15120	580	*1053*	829	1641
yale	*N*	*EUA*	15563	746	*827*	733	2280
rice	*N*	*EUA*	3687	651	*619*	537	1548
syracuse	*N*	*EUA*	2449	300	*451*	339	668
TOTAL			291375	30375	**35710**	27635	68860

Tabela D.2: Número de artigos por universidade e por área temática

Universidade	Brasileira	Todas	**ENG**	INTER	LIFE	MED	SCI	SOC
usp	**S**	16188	**2364**	686	5963	5089	5830	239
unicamp	**S**	7261	**1676**	201	2259	1795	3041	103
ufrj	S	6200	1517	225	2319	1206	2329	141
ufrgs	S	4167	787	237	1709	1067	1416	61
ufmg	S	3747	647	131	1397	1283	1214	90
puc-rio	S	991	483	19	89	58	513	33
ufpe	S	1721	448	42	542	258	798	28
unb	S	1743	311	62	641	280	788	55
texas-austin	**N**	42143	**5709**	3208	11996	18610	8277	2537
illinois	**N**	23584	**5411**	2141	6238	4911	7660	2221
berkeley	N	20097	4358	1531	4956	1670	9482	1784
mit	N	13852	4343	860	2229	1089	7227	907
imperial	N	16353	3765	695	4556	4510	5751	430
stanford	N	19782	3302	2056	5025	5937	5643	1510
cmellon	N	5575	2794	440	601	238	2205	583
waterloo	N	6332	2744	373	1178	544	2695	384
caltech	*N*	10565	*2162*	591	1144	172	7626	240
cornell	*N*	17253	*2146*	1417	6629	4343	4325	1290
northwestern	*N*	13092	*2027*	898	2752	4841	3583	1221
harvard	*N*	39728	*1800*	4789	13692	18951	4973	2715
princeton	*N*	8917	*1653*	648	1316	258	5228	838
duke	*N*	15120	*1252*	1644	4951	6823	2428	1035
rice	*N*	3687	*1021*	222	708	248	1918	322
yale	*N*	15563	*917*	2464	5087	5537	3206	1197
syracuse	*N*	2449	*454*	299	355	304	911	556
TOTAL		291375	**51117**	23715	82719	84351	87984	19199

Tabela D.3: Total de citações de artigos por universidade e por área de pesquisa

Universidade	Brasileira	País	Todas	CHEM	**CS**	MATH	PHYS
usp	**S**	**BRASIL**	46724	7575	**1688**	1518	13381
unicamp	**S**	**BRASIL**	18250	5446	**962**	751	5351
ufrj	S	BRASIL	15196	2077	833	734	3668
ufrgs	S	BRASIL	10931	1760	464	403	2509
ufmg	S	BRASIL	9084	1595	396	317	2678
ufpe	S	BRASIL	2813	719	326	277	970
puc-rio	S	BRASIL	1878	492	309	240	1092
unb	S	BRASIL	3357	639	224	181	1168
mit	**N**	**EUA**	141013	19524	**9752**	7002	51216
berkeley	**N**	**EUA**	150650	20624	**9510**	7441	56808
texas-austin	N	EUA	287575	15961	8554	5352	31744
stanford	N	EUA	165404	10393	7762	5364	28907
illinois	N	EUA	109842	15908	7229	4887	30581
harvard	N	EUA	407637	13758	6138	5459	23943
caltech	N	EUA	99975	10796	5073	3200	56492
princeton	N	EUA	70094	4395	4513	3822	40841
cornell	N	EUA	113807	7531	4036	2901	19345
cmellon	N	EUA	29975	4321	3940	3154	15496
imperial	N	INGLATERRA	107008	10884	3397	2547	24143
northwestern	N	EUA	91661	14872	3177	2533	19974
waterloo	N	CANADA	17113	3003	2913	2033	6339
duke	N	EUA	120958	4122	2455	1986	9912
yale	N	EUA	131712	6418	2020	1650	17080
rice	*N*	*EUA*	26429	6969	*1526*	1094	13028
syracuse	*N*	*EUA*	9683	1479	*626*	485	3977
TOTAL			1911926	180389	**79840**	58823	375482

Tabela D.4: Total de citações de artigos por universidade e por área temática

Universidade	Brasileira	Todas	**ENG**	INTER	LIFE	MED	SCI	SOC
usp	**S**	46724	**3920**	2805	16066	14340	18882	290
unicamp	**S**	18250	**2797**	339	5854	5419	8174	93
ufrj	S	15196	1861	809	6323	4288	5007	133
ufmg	S	9084	954	796	3377	2681	3427	89
ufrgs	S	10931	927	672	4236	3121	4106	83
puc-rio	S	1878	633	63	204	93	1230	56
ufpe	S	2813	525	29	888	536	1496	30
unb	S	3357	396	117	1608	383	1534	63
mit	**N**	141013	**19998**	26212	35579	11635	62874	3368
berkeley	**N**	150650	**17740**	23091	43806	11968	72746	3847
texas-austin	N	287575	15876	30697	93473	139790	44192	4667
illinois	N	109842	14307	13062	35265	23041	39776	3913
stanford	N	165404	11673	29779	53601	49849	38370	3997
harvard	N	407637	10431	73969	161565	172654	38025	8176
imperial	N	107008	9192	9405	34300	39042	31636	862
northwestern	N	91661	9064	7948	23121	36417	25905	3173
cmellon	N	29975	8774	3155	3812	1277	16716	1534
caltech	N	99975	8758	13356	13533	1100	70184	580
cornell	N	113807	6860	17743	45078	30393	26166	2841
rice	N	26429	6071	2393	4837	1556	16209	686
princeton	N	70094	5538	8772	11617	1783	45734	2364
waterloo	N	17113	4392	1345	4058	2283	8800	524
duke	N	120958	4053	18238	42731	55498	14603	3382
yale	*N*	131712	*3511*	28437	52442	40497	23534	2870
syracuse	*N*	9683	*1005*	1110	1752	1289	5313	1040
TOTAL		1911926	**154933**	271747	635579	584996	510269	43623

Tabela D.5: Número de artigos em periódicos por universidade e por área de pesquisa

Universidade	Brasileira	País	Todas	CHEM	**CS**	MATH	PHYS
usp	**S**	**BRASIL**	15959	2751	**1240**	1041	3975
unicamp	**S**	**BRASIL**	7108	1788	**817**	589	2273
ufrj	S	BRASIL	6002	1042	678	558	1680
ufmg	S	BRASIL	3654	641	356	263	882
ufrgs	S	BRASIL	4043	611	281	227	973
ufpe	S	BRASIL	1606	379	252	206	592
puc-rio	S	BRASIL	901	220	242	171	477
unb	S	BRASIL	1704	300	228	186	626
texas-austin	**N**	**EUA**	41448	2757	**3756**	2618	6434
illinois	**N**	**EUA**	22942	2763	**3432**	2496	6363
berkeley	N	EUA	19512	2624	2552	1887	7483
mit	N	EUA	13323	2029	2544	1730	6342
stanford	N	EUA	19372	1373	2215	1507	4355
waterloo	N	CANADA	5969	785	1879	1331	2365
harvard	N	EUA	39331	1435	1562	1347	3157
cmellon	N	EUA	4995	727	1556	1214	2214
princeton	N	EUA	8769	846	1497	1256	4659
imperial	N	INGLATERRA	15801	2063	1493	1083	4541
caltech	N	EUA	10340	1227	1418	850	6187
cornell	N	EUA	17005	1280	1396	1047	3220
northwestern	*N*	*EUA*	12976	1714	*1059*	882	3082
duke	*N*	*EUA*	14990	580	*980*	756	1640
yale	*N*	*EUA*	15383	743	*737*	643	2268
rice	*N*	*EUA*	3579	647	*535*	453	1533
syracuse	*N*	*EUA*	2403	299	*408*	296	666
TOTAL			284682	30018	**31027**	22953	68206

Tabela D.6: Número de artigos em periódicos por universidade e por *Broad Subject Fields*

Universidade	Brasileira	Todas	**ENG**	INTER	LIFE	MED	SCI	SOC
usp	**S**	15959	**2165**	667	5962	5088	5817	238
unicamp	**S**	7108	**1537**	197	2259	1795	3029	103
ufrj	S	6002	1339	221	2319	1202	2316	140
ufrgs	S	4043	667	236	1709	1067	1411	61
ufmg	S	3654	562	123	1397	1283	1213	90
puc-rio	S	901	395	19	89	58	512	32
ufpe	S	1606	333	42	542	258	798	28
unb	S	1704	273	62	641	280	787	55
texas-austin	**N**	41448	**5175**	3117	11992	18574	8248	2514
illinois	**N**	22942	**4889**	2106	6237	4899	7621	2180
berkeley	N	19512	3865	1505	4954	1666	9381	1770
mit	N	13323	3852	848	2224	1083	7190	900
imperial	N	15801	3283	673	4551	4488	5705	430
stanford	N	19372	2968	2016	5023	5928	5625	1492
waterloo	N	5969	2385	370	1173	544	2694	383
cmellon	N	4995	2234	429	600	238	2198	578
caltech	*N*	10340	*2016*	585	1143	171	7510	240
cornell	*N*	17005	*1975*	1378	6627	4324	4300	1285
northwestern	*N*	12976	*1961*	879	2751	4836	3575	1199
harvard	*N*	39331	*1582*	4679	13681	18913	4963	2697
princeton	*N*	8769	*1545*	628	1316	258	5210	833
duke	*N*	14990	*1177*	1615	4950	6811	2426	1023
rice	*N*	3579	*928*	221	706	247	1906	316
yale	*N*	15383	*826*	2412	5087	5520	3195	1187
syracuse	*N*	2403	*409*	299	355	304	910	552
TOTAL		284682	**45612**	23192	82678	84171	87483	19019

Tabela D.7: Total de citações de artigos em periódicos por universidade e por área de pesquisa

Universidade	Brasileira	País	Todas	CHEM	**CS**	MATH	PHYS
usp	**S**	**BRASIL**	46591	7555	**1597**	1427	13360
unicamp	**S**	**BRASIL**	18195	5419	**934**	723	5335
ufrj	S	BRASIL	15114	2063	779	680	3658
ufrgs	S	BRASIL	10887	1757	425	364	2506
ufmg	S	BRASIL	9038	1571	387	308	2675
ufpe	S	BRASIL	2785	718	299	250	969
puc-rio	S	BRASIL	1843	490	276	207	1091
unb	S	BRASIL	3347	639	214	171	1168
mit	**N**	**EUA**	140545	19498	**9435**	6685	51147
berkeley	**N**	**EUA**	150061	20608	**9196**	7128	56677
texas-austin	N	EUA	286861	15953	8413	5211	31707
stanford	N	EUA	164580	10390	7178	4780	28895
illinois	N	EUA	109378	15901	6955	4613	30560
harvard	N	EUA	406806	13756	6006	5327	23888
caltech	N	EUA	99625	10791	4997	3124	56357
princeton	N	EUA	70010	4395	4451	3760	40838
cornell	N	EUA	113448	7527	3937	2802	19326
cmellon	N	EUA	29474	4252	3546	2760	15448
northwestern	N	EUA	91417	14866	3167	2523	19968
imperial	N	INGLATERRA	106519	10839	3139	2289	24078
waterloo	N	CANADA	16916	3002	2721	1841	6338
duke	N	EUA	120725	4122	2405	1936	9911
yale	N	EUA	131300	6417	1932	1562	17070
rice	*N*	*EUA*	26322	6966	*1459*	1027	13010
syracuse	*N*	*EUA*	9660	1476	*611*	470	3969
TOTAL			1905021	180110	**76665**	55649	374899

Tabela D.8: Total de citações de artigos por universidade e por *Broad Subject Fields*

Universidade	Brasileira	Todas	**ENG**	INTER	LIFE	MED	SCI	SOC
usp	S	46591	**3810**	2784	16066	14340	18878	290
unicamp	S	18195	**2753**	339	5854	5419	8163	93
ufrj	S	15114	1800	800	6323	4286	4999	130
ufmg	S	9038	921	783	3377	2681	3427	89
ufrgs	S	10887	885	670	4236	3121	4105	83
puc-rio	S	1843	598	63	204	93	1230	56
ufpe	S	2785	497	29	888	536	1496	30
unb	S	3347	386	117	1608	383	1534	63
mit	N	140545	**19612**	26169	35565	11621	62824	3362
berkeley	N	150061	**17285**	22986	43805	11961	72603	3844
texas-austin	N	286861	15695	30233	93463	139726	44162	4658
illinois	N	109378	14014	12933	35264	23022	39745	3911
stanford	N	164580	11077	29567	53597	49846	38361	3975
harvard	N	406806	10253	73488	161510	172565	37970	8125
northwestern	N	91417	9045	7750	23109	36409	25899	3155
imperial	N	106519	8869	9318	34259	39025	31580	862
caltech	N	99625	8552	13327	13533	1100	69940	580
cmellon	N	29474	8332	3147	3811	1277	16692	1505
cornell	N	113448	6743	17517	45077	30380	26151	2838
rice	N	26322	5988	2390	4817	1556	16194	686
princeton	N	70010	5471	8759	11617	1783	45731	2361
waterloo	N	16916	4196	1344	4044	2283	8800	524
duke	N	120725	4002	18086	42731	55477	14602	3373
yale	*N*	131300	*3415*	28229	52442	40393	23525	2867
syracuse	*N*	9660	*982*	1110	1752	1289	5308	1040
TOTAL		1905021	**151162**	269432	635411	584649	509653	43475

Apêndice E

Estatísticas de artigos dos departamentos/institutos de Computação

Neste apêndice, as estatísticas de artigos produzidos pelos departamentos e institutos de computação e engenharia da computação das universidades escolhidas para o experimento e estes dados estão na forma de tabelas. Estas tabelas foram agrupadas por grupos de categorias de pesquisa de periódicos 3.2, que são baseados nas definições de área de pesquisa afins (divididos em: CS, CHEM, PHYS e MATH) que foram identificados via consulta a especialistas de cada área. As tabelas também foram agrupadas em outra definição destes grupos chamada de *Broad Subject Fields* ou áreas temáticas (divisos em ENG, INTER, LIFE, MED, SCI e SOC) que é utilizada pelo ranking de universidades ARWU [12].

Totalizou-se a quantidade de artigos e citações por departamento/instituto versus áreas de pesquisa e por departamento/instituto versus *Broad Subject Fields*. Tabelas separadas destacam a quantidade de artigos produzidos especificamente em periódicos, o que nos permite observar que nos dados utilizados há poucas ocorrências de artigos publicados que não sejam de periódicos em relação ao total de artigos.

Nos relatórios agrupados por áreas de pesquisa (CS, CHEM, PHYS e MATH), os dados foram ordenados de forma a destacar a importância das instituições brasileiras em produção científica (total de artigos e total de citações) especificamente do total da produção de cada departamento ou instituto de ciência da computação ou engenharia da computação em cada universidade. As instituições que não são brasileiras também estão ordenadas desta forma, mas em separado e se encontram logo abaixo das universidades brasileiras. Para destacar estes critérios de classificação, foi colocada uma coluna identificando o país da instituição. Pelo mesmo motivo o título da coluna **Todas** está em negrito

em nas tabelas agrupadas por área de pesquisa e **Todos** nas tabelas agrupadas por áreas temáticas.

Tabela E.1: Citações de artigos de professores de departamentos/inst. de computação por área temática

Universidade	Brasileira	**Todos**	ENG	INTER	LIFE	MED	SCI	SOC	Outros
ufrj	S	103	62	0	1	1	46	5	1
unicamp	S	102	93	0	6	11	15	6	0
ufrgs	S	85	72	0	0	0	22	5	0
ufpe	S	48	48	0	0	0	0	0	0
ufmg	S	46	46	0	0	2	2	6	0
puc-rio	S	36	36	0	0	2	0	0	0
usp	S	36	36	0	0	0	0	0	0
mit	N	6537	2692	1512	729	176	2497	61	0
yale	N	2462	232	1266	988	1	105	8	0
cmellon	N	2017	762	8	81	27	1270	22	0
berkeley	N	1528	978	340	198	6	237	16	0
stanford	N	1411	983	8	335	3	259	1	0
illinois	N	1364	1232	4	57	1	271	91	0
northwestern	N	1190	602	4	14	35	752	26	0
rice	N	962	340	8	29	0	796	0	0
duke	N	832	272	371	32	0	334	2	0
caltech	N	740	585	18	260	0	372	8	0
waterloo	N	695	537	0	69	4	280	3	0
princeton	N	586	346	2	146	0	242	38	0
texas-austin	N	543	386	7	98	2	204	3	0
imperial	N	423	399	0	14	126	28	2	0
cornell	N	423	364	23	9	2	48	18	0
harvard	N	158	152	8	3	1	22	10	0
syracuse	N	92	92	0	0	0	5	0	0
Total		21873	10934	3579	3056	397	7642	331	1

Tabela E.2: Citações de artigos de professores de departamentos/inst. de computação por área de pesquisa

Universidade	País	**Todas**	CHEM	CS	MATH	PHYS	Outras
ufrj	BRASIL	103	0	102	99	48	1
unicamp	BRASIL	102	2	98	98	36	2
ufrgs	BRASIL	85	1	77	71	35	1
ufpe	BRASIL	48	0	48	43	13	0
ufmg	BRASIL	46	0	46	41	7	0
puc-rio	BRASIL	36	0	36	36	7	0
usp	BRASIL	36	0	36	33	14	0
mit	EUA	6537	448	2542	1477	3736	2199
yale	EUA	2462	7	254	252	83	2208
cmellon	EUA	2017	8	779	661	1418	43
berkeley	EUA	1528	0	1068	963	397	460
stanford	EUA	1411	0	1124	1040	262	287
illinois	EUA	1364	4	1288	1047	703	58
northwestern	EUA	1190	13	602	255	1048	43
rice	EUA	962	716	203	148	341	14
duke	EUA	832	238	221	220	158	372
caltech	EUA	740	2	618	590	295	75
waterloo	CANADA	695	82	591	547	303	14
princeton	EUA	586	0	492	461	228	94
texas-austin	EUA	543	0	418	412	234	42
imperial	INGLATERRA	423	6	406	390	186	14
cornell	EUA	423	0	383	344	177	40
harvard	EUA	158	0	152	152	64	6
syracuse	EUA	92	0	92	34	72	0
Total		21873	1527	11230	8990	9573	5970

Tabela E.3: Número de artigos de professores de departamentos/inst. de computação por área de pesquisa

Universidade	País	**Todas**	CHEM	CS	MATH	PHYS	Outras
ufrj	BRASIL	202	0	197	195	52	2
ufrgs	BRASIL	150	3	140	121	38	5
unicamp	BRASIL	133	1	127	124	37	5
ufpe	BRASIL	104	0	102	98	9	2
usp	BRASIL	79	0	74	71	12	4
ufmg	BRASIL	75	0	74	61	15	1
puc-rio	BRASIL	74	0	74	74	6	0
mit	EUA	875	50	669	401	550	50
illinois	EUA	812	4	770	673	302	14
waterloo	CANADA	469	2	457	426	177	6
cmellon	EUA	455	2	403	386	96	13
northwestern	EUA	404	7	304	209	282	15
stanford	EUA	393	0	373	340	87	20
berkeley	EUA	353	0	334	313	83	18
imperial	INGLATERRA	342	3	331	325	56	10
texas-austin	EUA	326	0	298	294	95	12
rice	EUA	258	58	186	156	85	13
cornell	EUA	237	1	219	196	65	17
yale	EUA	229	3	143	137	57	86
princeton	EUA	225	0	204	193	73	19
duke	EUA	130	7	113	110	46	9
caltech	EUA	130	1	120	108	65	5
harvard	EUA	95	0	93	92	23	2
syracuse	EUA	90	1	88	26	73	0
Total		6456	143	5723	4962	2315	325

Tabela E.4: Número de artigos de professores de departamentos/inst. de computação por *Broad Subject Fields*

Universidade	Brasileira	**Todos**	ENG	INTER	LIFE	MED	SCI	SOC	Outros
ufrj	S	202	160	0	4	1	49	10	1
ufrgs	S	150	142	1	2	1	19	4	0
unicamp	S	133	110	0	2	5	26	9	0
ufpe	S	104	100	0	2	0	3	0	0
usp	S	79	74	0	3	2	2	0	0
ufmg	S	75	74	0	1	2	3	9	0
puc-rio	S	74	71	2	0	1	5	0	0
mit	N	875	670	22	39	18	358	21	1
illinois	N	812	741	12	12	3	149	37	0
waterloo	N	469	401	0	22	3	123	8	0
cmellon	N	455	396	6	19	5	72	11	0
northwestern	N	404	299	3	5	12	149	10	0
stanford	N	393	346	4	21	2	59	5	0
berkeley	N	353	302	7	29	3	77	4	0
imperial	N	342	325	2	10	24	19	5	0
texas-austin	N	326	285	4	17	5	61	2	0
rice	N	258	201	5	14	1	81	2	0
cornell	N	237	214	7	6	5	22	9	1
yale	N	229	118	18	82	5	52	5	0
princeton	N	225	178	3	24	0	59	13	0
duke	N	130	108	4	13	2	50	2	0
caltech	N	130	99	2	5	0	39	1	0
harvard	N	95	92	3	3	1	10	3	0
syracuse	N	90	88	0	0	0	11	0	0
Total		6456	5440	104	330	99	1451	168	3

Tabela E.5: Total de citações de artigos em periódicos por departamento de computação e por área de pesquisa

Universidade	País	**Todas**	CHEM	CS	MATH	PHYS	Outras
unicamp	BRASIL	89	2	85	85	36	2
ufrj	BRASIL	78	0	77	74	48	1
ufrgs	BRASIL	54	1	46	40	35	1
ufmg	BRASIL	39	0	39	34	7	0
ufpe	BRASIL	30	0	30	25	13	0
usp	BRASIL	24	0	24	21	14	0
puc-rio	BRASIL	22	0	22	22	7	0
mit	EUA	6418	448	2423	1358	3736	2199
yale	EUA	2438	7	230	228	83	2208
cmellon	EUA	1869	8	631	513	1418	43
berkeley	EUA	1456	0	996	891	397	460
illinois	EUA	1185	4	1109	868	703	58
northwestern	EUA	1183	13	595	248	1048	43
stanford	EUA	1094	0	807	723	262	287
rice	EUA	897	716	141	86	341	11
duke	EUA	795	238	184	183	158	372
caltech	EUA	700	2	578	550	295	75
waterloo	CANADA	582	82	478	434	303	14
princeton	EUA	579	0	485	454	228	94
texas-austin	EUA	491	0	366	360	234	42
cornell	EUA	383	0	343	304	177	40
imperial	INGLATERRA	272	3	258	242	183	14
harvard	EUA	122	0	116	116	64	6
syracuse	EUA	89	0	89	31	72	0
Total		20374	1524	9737	7497	9570	5967

Tabela E.6: Total de citações de artigos em periódicos por departamento de computação e por *Broad Subject Fields*

Universidade	Brasileira	**Todos**	ENG	INTER	LIFE	MED	SCI	SOC	Outros
unicamp	BRASIL	89	80	0	6	11	15	6	0
ufrj	BRASIL	78	37	0	1	1	46	5	1
ufrgs	BRASIL	54	41	0	0	0	22	5	0
ufmg	BRASIL	39	39	0	0	2	2	6	0
ufpe	BRASIL	30	30	0	0	0	0	0	0
usp	BRASIL	24	24	0	0	0	0	0	0
puc-rio	BRASIL	22	22	0	0	2	0	0	0
mit	EUA	6418	2573	1512	729	176	2497	61	0
yale	EUA	2438	208	1266	988	1	105	8	0
cmellon	EUA	1869	614	8	81	27	1270	22	0
berkeley	EUA	1456	906	340	198	6	237	16	0
illinois	EUA	1185	1053	4	57	1	271	91	0
northwestern	EUA	1183	595	4	14	35	752	26	0
stanford	EUA	1094	666	8	335	3	259	1	0
rice	EUA	897	278	5	29	0	796	0	0
duke	EUA	795	235	371	32	0	334	2	0
caltech	EUA	700	545	18	260	0	372	8	0
waterloo	CANADA	582	424	0	57	4	280	3	0
princeton	EUA	579	339	2	146	0	242	38	0
texas-austin	EUA	491	334	7	97	2	204	3	0
cornell	EUA	383	324	23	9	2	48	18	0
imperial	INGLATERRA	272	248	0	14	126	28	2	0
harvard	EUA	122	116	8	3	1	22	10	0
syracuse	EUA	89	89	0	0	0	5	0	0
Total		20374	9438	3576	3043	397	7642	331	1

Tabela E.7: Número de artigos em periódicos por departamento de computação e por área de pesquisa

Universidade	País	Todas	CHEM	CS	MATH	PHYS	Outras
ufrj	BRASIL	108	0	103	101	52	2
unicamp	BRASIL	82	1	76	73	37	5
ufrgs	BRASIL	72	2	63	44	37	5
ufmg	BRASIL	49	0	48	35	15	1
usp	BRASIL	38	0	33	30	12	4
ufpe	BRASIL	31	0	29	25	9	2
puc-rio	BRASIL	29	0	29	29	6	0
mit	EUA	752	50	546	278	550	50
illinois	EUA	571	4	529	432	302	14
northwestern	EUA	366	7	266	171	282	15
waterloo	CANADA	325	2	313	282	177	6
cmellon	EUA	290	2	239	222	96	12
stanford	EUA	277	0	257	224	87	20
berkeley	EUA	248	0	229	208	83	18
texas-austin	EUA	212	0	184	180	95	12
yale	EUA	200	3	114	108	57	86
princeton	EUA	186	0	166	155	73	18
rice	EUA	186	58	115	85	85	12
cornell	EUA	171	1	153	130	65	17
imperial	INGLATERRA	135	2	125	119	55	10
caltech	EUA	102	1	92	80	64	5
duke	EUA	86	7	69	66	46	9
syracuse	EUA	84	1	82	20	73	0
harvard	EUA	60	0	58	57	23	2
Total		4517	141	3788	3027	2312	323

Tabela E.8: Número de artigos em periódicos por departamento de computação e por *Broad Subject Fields*

Universidade	Brasileira	**Todos**	ENG	INTER	LIFE	MED	SCI	SOC	Outros
ufrj	S	108	66	0	4	1	49	10	1
unicamp	S	82	59	0	2	5	26	9	0
ufrgs	S	72	64	1	2	1	19	4	0
ufmg	S	49	48	0	1	2	3	9	0
usp	S	38	33	0	3	2	2	0	0
ufpe	S	31	27	0	2	0	3	0	0
puc-rio	S	29	26	2	0	1	5	0	0
mit	N	752	547	22	39	18	358	21	1
illinois	N	571	500	12	12	3	149	37	0
northwestern	N	366	261	3	5	12	149	10	0
waterloo	N	325	257	0	18	3	123	8	0
cmellon	N	290	232	5	19	5	72	11	0
stanford	N	277	230	4	20	2	59	5	0
berkeley	N	248	197	7	29	3	77	4	0
texas-austin	N	212	171	4	16	5	61	2	0
yale	N	200	89	18	82	5	52	5	0
princeton	N	186	140	2	24	0	59	13	0
rice	N	186	130	4	14	1	81	2	0
cornell	N	171	148	7	6	5	22	9	1
imperial	N	135	118	2	10	24	19	5	0
caltech	N	102	72	2	5	0	38	1	0
duke	N	86	64	4	13	2	50	2	0
syracuse	N	84	82	0	0	0	11	0	0
harvard	N	60	57	3	3	1	10	3	0
Total		4517	3504	102	324	99	1450	168	3

Tabela E.9: Proporção em quantidade e em citações dos artigos produzidos pelo professores de departamentos/inst. de computação que são da área CS

		Artigos por área		Citações por área		% da área de CS	
Universidade	País	Todas	CS	Todas	CS	Artigos	Citações
ufrj	BRASIL	202	197	103	102	97,52	99,03
ufrgs	BRASIL	150	140	85	77	93,33	90,59
unicamp	BRASIL	133	127	102	98	95,49	96,08
ufpe	BRASIL	104	102	48	48	98,08	100,00
usp	BRASIL	79	74	36	36	93,67	100,00
ufmg	BRASIL	75	74	40	40	98,67	100,00
puc-rio	BRASIL	74	74	36	36	100,00	100,00
mit	EUA	875	669	6537	2542	76,46	38,89
illinois	EUA	812	770	1364	1288	94,83	94,43
waterloo	CANADA	469	457	695	591	97,44	85,04
cmellon	EUA	455	403	2017	779	88,57	38,62
northwestern	EUA	404	304	1190	602	75,25	50,59
stanford	EUA	393	373	1411	1124	94,91	79,66
berkeley	EUA	353	334	1528	1068	94,62	69,90
imperial	INGLATERRA	342	331	423	406	96,78	95,98
texas-austin	EUA	326	298	543	418	91,41	76,98
rice	EUA	258	186	962	203	72,09	21,10
cornell	EUA	237	219	423	383	92,41	90,54
yale	EUA	229	143	2462	254	62,45	10,32
princeton	EUA	225	204	586	492	90,67	83,96
duke	EUA	130	113	832	221	86,92	26,56
caltech	EUA	130	120	740	618	92,31	83,51
harvard	EUA	95	93	158	152	97,89	96,20
syracuse	EUA	90	88	92	92	97,78	100,00
Total		6456	5723	21873	11230	88,65	51,34

Tabela E.10: Proporção em quantidade e em citações dos artigos produzidos pelo professores de departamentos/inst. de computação que são de área temática ENG

		Artigos		Citações		% de ENG	
Universidade	País	Todos	ENG	Todos	ENG	Artigos	Citações
ufrj	BRASIL	202	160	103	62	79,21	60,19
ufrgs	BRASIL	150	142	85	72	94,67	84,71
unicamp	BRASIL	133	110	102	93	82,71	91,18
ufpe	BRASIL	104	100	48	48	96,15	100,00
usp	BRASIL	79	74	36	36	93,67	100,00
ufmg	BRASIL	75	74	40	40	98,67	100,00
puc-rio	BRASIL	74	71	36	36	95,95	100,00
mit	EUA	875	670	6537	2692	76,57	41,18
illinois	EUA	812	741	1364	1232	91,26	90,32
waterloo	CANADA	469	401	695	537	85,50	77,27
cmellon	EUA	455	396	2017	762	87,03	37,78
northwestern	EUA	404	299	1190	602	74,01	50,59
stanford	EUA	393	346	1411	983	88,04	69,67
berkeley	EUA	353	302	1528	978	85,55	64,01
imperial	INGLATERRA	342	325	423	399	95,03	94,33
texas-austin	EUA	326	285	543	386	87,42	71,09
rice	EUA	258	201	962	340	77,91	35,34
cornell	EUA	237	214	423	364	90,3	86,05
yale	EUA	229	118	2462	232	51,53	9,42
princeton	EUA	225	178	586	346	79,11	59,04
duke	EUA	130	108	832	272	83,08	32,69
caltech	EUA	130	99	740	585	76,15	79,05
harvard	EUA	95	92	158	152	96,84	96,20
syracuse	EUA	90	88	92	92	97,78	100,00
Total		6456	5440	21873	10934	84,26	49,99

Tabela E.11: Proporção em quantidade e em citações dos artigos em periódicos produzidos pelo professores de departamentos/inst. de computação que são da área CS

		Artigos por área		Citações por área		% da área de CS	
Universidade	País	Todas	CS	Todas	CS	Artigos	Citações
ufrj	BRASIL	108	103	78	77	95,37	98,72
unicamp	BRASIL	82	76	89	85	92,68	95,51
ufrgs	BRASIL	72	63	54	46	87,50	85,19
ufmg	BRASIL	49	48	33	33	97,96	100,00
usp	BRASIL	38	33	24	24	86,84	100,00
ufpe	BRASIL	31	29	30	30	93,55	100,00
puc-rio	BRASIL	29	29	22	22	100,00	100,00
mit	EUA	752	546	6418	2423	72,61	37,75
illinois	EUA	571	529	1185	1109	92,64	93,59
northwestern	EUA	366	266	1183	595	72,68	50,30
waterloo	CANADA	325	313	582	478	96,31	82,13
cmellon	EUA	290	239	1869	631	82,41	33,76
stanford	EUA	277	257	1094	807	92,78	73,77
berkeley	EUA	248	229	1456	996	92,34	68,41
texas-austin	EUA	212	184	491	366	86,79	74,54
yale	EUA	200	114	2438	230	57,00	9,43
princeton	EUA	186	166	579	485	89,25	83,77
rice	EUA	186	115	897	141	61,83	15,72
cornell	EUA	171	153	383	343	89,47	89,56
imperial	INGLATERRA	135	125	272	258	92,59	94,85
caltech	EUA	102	92	700	578	90,20	82,57
duke	EUA	86	69	795	184	80,23	23,14
syracuse	EUA	84	82	89	89	97,62	100,00
harvard	EUA	60	58	122	116	96,67	95,08
Total		4517	3788	20374	9737	83,86	47,79

Tabela E.12: Proporção em quantidade e em citações dos artigos em periódicos produzidos pelo professores de departamentos/inst. de computação que são de área temática ENG

		Artigos		Citações		% de ENG	
Universidade	País	Todos	ENG	Todos	ENG	Artigos	Citações
ufrj	BRASIL	108	66	78	37	61,11	47,44
unicamp	BRASIL	82	59	89	80	71,95	89,89
ufrgs	BRASIL	72	64	54	41	88,89	75,93
ufmg	BRASIL	49	48	33	33	97,96	100,00
usp	BRASIL	38	33	24	24	86,84	100,00
ufpe	BRASIL	31	27	30	30	87,10	100,00
puc-rio	BRASIL	29	26	22	22	89,66	100,00
mit	EUA	752	547	6418	2573	72,74	40,09
illinois	EUA	571	500	1185	1053	87,57	88,86
northwestern	EUA	366	261	1183	595	71,31	50,30
waterloo	CANADA	325	257	582	424	79,08	72,85
cmellon	EUA	290	232	1869	614	80,00	32,85
stanford	EUA	277	230	1094	666	83,03	60,88
berkeley	EUA	248	197	1456	906	79,44	62,23
texas-austin	EUA	212	171	491	334	80,66	68,02
yale	EUA	200	89	2438	208	44,50	8,53
princeton	EUA	186	140	579	339	75,27	58,55
rice	EUA	186	130	897	278	69,89	30,99
cornell	EUA	171	148	383	324	86,55	84,60
imperial	INGLATERRA	135	118	272	248	87,41	91,18
caltech	EUA	102	72	700	545	70,59	77,86
duke	EUA	86	64	795	235	74,42	29,56
syracuse	EUA	84	82	89	89	97,62	100,00
harvard	EUA	60	57	122	116	95,00	95,08
Total		4517	3504	20374	9438	77,57	46,32

Apêndice F

Ranking de Prestígio na rede de coautoria

Na Tabela F.1 são apresentados os resultados da pontuação do prestígio dos autores no principal componente da rede de coautoria obtido no experimento. Os autores estão identificados pelo seu nome completo conforme a lista oferecida ao sistema do quadro de pesquisadores de cada instituição. Para os pesquisadores que não pertencem ao quadro destas instituições, o conteúdo da coluna universidade está vazio, seu nome está conforme o nome encontrado em uma de suas publicações e está em caixa baixa. A listagem é apresentada em ordem alfabética, pois não há interesse em destacar o aspecto do prestígio individual dos pesquisadores, e sim utilizar a informação para agregar prestígio a um contexto maior como a instituição a qual são afiliados. No entanto, se for necessário reproduzir os resultados o dado da posição no *ranking* é apresentado conforme foi calculado no experimento.

Tabela F.1: Prestígio dos autores na rede de coautoria

Autor	Universidade	Posição
Adam Finkelstein	princeton	153
Alan S Willsky	mit	181
Alex Lopez-ortiz	waterloo	90
Alfred Menezes	waterloo	96
Allen Taflove	northwestern	163
Alok Choudhary	northwestern	36
Aloysius K Mok	texas-austin	86
Ana Lucia Cetertich Bazzan	ufrgs	187
Anantha P Chandrakasan	mit	59

Andrew W Appel	princeton	171
Anupam Gupta	cmellon	21
Avrim Blum	cmellon	62
Balaji Prabhakar	stanford	28
Benjamin Wan-sang Wah	illinois	177
Bernard Chazelle	princeton	56
Bernd Sturmfels	berkeley	17
Berthier Ribeiro Neto	ufmg	190
Biao Chen	syracuse	137
Carlos Jose Pereira De Lucena	puc-rio	49
Celina Miraglia Herrera De Figueiredo	ufrj	37
Chandrajit Bajaj	texas-austin	20
Christos Faloutsos	cmellon	45
Christos H Papadimitriou	berkeley	72
Chung-chieh Lee	northwestern	118
Claudia Maria Bauzer Medeiros	unicamp	165
Claudia Maria Lima Werner	ufrj	162
Daphne Koller	stanford	131
David A Padua	illinois	110
David E Goldberg	illinois	13
David I August	princeton	113
David K Gifford	mit	154
David L Dill	stanford	182
David Perreault	mit	166
David R Karger	mit	40
David Walker	princeton	123
Dimitri A Antoniadis	mit	19
Douglas R Stinson	waterloo	67
E Allen Emerson	texas-austin	129
Edmundo Roberto Mauro Madeira	unicamp	54
Eduardo Sany Laber	puc-rio	161
Erich P Ippen	mit	7
Frank Pfenning	cmellon	79
Franz X Kartner	mit	43
Fredo Durand	mit	168
Gene H Golub	stanford	23

George C Necula	berkeley	68
George Labahn	waterloo	150
Grigore Rosu	illinois	15
Gul A Agha	illinois	60
Guy E Blelloch	cmellon	167
Hai Zhou	northwestern	87
Hari Balakrishnan	mit	84
Hector Garcia-molina	stanford	24
Henry I Smith	mit	9
Herbert Edelsbrunner	duke	89
Hong Wang	syracuse	159
Horace P Yuen	northwestern	186
Hui Zhang	cmellon	85
Ion Stoica	berkeley	18
J Ian Munro	waterloo	82
j. lee		139
j. li		185
Jacob K White	mit	170
James Aspnes	yale	164
James C Browne	texas-austin	103
James G Fujimoto	mit	112
James W Demmel	berkeley	126
Jano Moreira De Souza	ufrj	14
Jayme Luiz Szwarcfiter	ufrj	100
Jean Ponce	illinois	136
Jeff Erickson	illinois	158
Jeffrey H Lang	mit	25
Jeffrey H Shapiro	mit	145
Jeffrey Shallit	waterloo	39
Jennifer C Hou	illinois	26
Jennifer Rexford	princeton	102
Jennifer Widom	stanford	130
Jessica K Hodgins	cmellon	116
Jiawei Han	illinois	10
Jin A Kong	mit	8
Jinbo Xu	waterloo	114

Joan Feigenbaum	yale	58
Joel Mambretti	northwestern	75
Johannes E Gehrke	cornell	94
John C Doyle	caltech	47
John D Kubiatowicz	berkeley	172
John H Reif	duke	122
John Mellor-crummey	rice	69
Jon Kleinberg	cornell	78
Jorge Stolfi	unicamp	174
Jose Meseguer	illinois	1
Josep Torrellas	illinois	144
Joseph Y Halpern	cornell	111
Judy L Hoyt	mit	142
Jun Yang	duke	109
Kai Li	princeton	175
Karl Crary	cmellon	189
Karl K Berggren	mit	157
Kei-hoi Cheung	yale	66
Keith D Cooper	rice	55
Kevin Chen Chuan Chang	illinois	179
Klara Nahrstedt	illinois	11
Krzysztof Czarnecki	waterloo	184
Larry Peterson	princeton	73
Lars Arge	duke	101
Laxmikant V Kale	illinois	133
Leonard Schulman	caltech	119
Leonidas J Guibas	stanford	12
Les Gasser	illinois	183
Lior Pachter	berkeley	134
Luay K Nakhleh	rice	143
Luca Trevisan	berkeley	115
Lui Sha	illinois	108
Luigi Carro	ufrgs	29
Lydia E Kavraki	rice	63
Madhu Sudan	mit	180
Mahesh Viswanathan	illinois	70

Manuela M Veloso	cmellon	3
Maria Das Gracas Volpe Nunes	usp	173
Marianne Winslett	illinois	138
Marta Lima De Queiros Mattoso	ufrj	51
Martin A Schmidt	mit	121
Martin H Schultz	yale	83
Michael I Jordan	berkeley	16
Michael K Reiter	cmellon	48
Michael L Honig	northwestern	35
Ming Li	waterloo	65
Ming-yang Kao	northwestern	53
Mohamed G Gouda	texas-austin	74
Moses Charikar	princeton	93
Muriel Medard	mit	44
Nancy A Lynch	mit	191
Narendra Ahuja	illinois	71
Nick Mckeown	stanford	42
Nikolaos V Sahinidis	illinois	95
Nitin H Vaidya	illinois	155
Olga G Troyanskaya	princeton	81
P Gardner	imperial	178
P R Kumar	illinois	124
Pankaj K Agarwal	duke	4
Peter Druschel	rice	120
Peter Stone	texas-austin	34
Pin-han Ho	waterloo	61
Pramod Varshney	syracuse	64
Prem Kumar	northwestern	41
Randy H Katz	berkeley	117
Raouf Boutaba	waterloo	2
Ras Bodik	berkeley	128
Ravi K Iyer	illinois	52
Ravindran Kannan	yale	146
Reid G Simmons	cmellon	50
Ricardo Da Silva Torres	unicamp	152
Risto Miikkulainen	texas-austin	140

Robert E Schapire	princeton	106
Robert T Morris	mit	132
Robin Cohen	waterloo	188
Roy H Campbell	illinois	88
Salil P Vadhan	harvard	91
Sanjoy K Mitter	mit	127
Sariel Har-peled	illinois	27
Satish Rao	berkeley	107
Scott Shenker	berkeley	33
Sebastian Thrun	stanford	32
Seda Ogrenci Memik	northwestern	160
Seng-tiong Ho	northwestern	135
Serafim Batzoglou	stanford	92
Sheldon Howard Jacobson	illinois	22
Steven Low	caltech	76
Szymon Rusinkiewicz	princeton	148
Takeo Kanade	cmellon	30
Tandy Warnow	texas-austin	80
Tapan Sarkar	syracuse	38
Terry P Orlando	mit	156
Therese Biedl	waterloo	57
Thomas Funkhouser	princeton	149
Thomas Labean	duke	169
Thomas S Huang	illinois	5
Timothy M Chan	waterloo	98
Tommi S Jaakkola	mit	99
Tuomas Sandholm	cmellon	151
Vitor Manuel De Morais Santos Costa	ufrj	104
Vladimir Bulovic	mit	77
W Luk	imperial	6
Xu Li	northwestern	125
y. x. liu		105
Yang Richard Yang	yale	31
Yehea I Ismail	northwestern	147
Yizhou Yu	illinois	176
Yuanyuan Y Zhou	illinois	97

Zohar Manna	stanford	141
Zygmunt J Haas	cornell	46

Apêndice G

Especificação das consultas de extração de artigos da WOS ISI

A ferramenta de análise de redes sociais em dados bibliográfico recebe como entrada arquivos obtidos via consulta a ferramenta WOS da ISI. Cada arquivo deve conter dados de artigos referentes à produção científica de uma dada universidade e na tabela abaixo são descritas as *strings* de consulta utilizada para obter esta informação por universidade:

Tabela G.1: *Strings* de consulta no formato WOS ISI por universidade

Universidade	*String* de consulta
berkeley	OG=(UNIV* CALIF* BERKELEY)
caltech	OG=(calif* inst* tech* or caltech) or SG=(calif* inst* tech* or caltech)
cmellon	og=carnegie mellon univ or sg=carnegie mellon univ
cornell	OG=(cornell univ) or sg=(cornell univ)
duke	OG=duke univ or sg=duke univ
harvard	OG=harvard univ
illinois	OG=univ illinois or SG= univ illinois
imperial	OG=imperial col*
mit	OG=(mit or mass* inst* tech*) or SG=(mit or mass* inst* tech*)
northwestern	OG=northwestern univ
princeton	OG=princeton univ or SG=princeton univ
puc-rio	(OG=(puc or pont* univ* cat* or cat* univ* or pont* cat* univ*) or SG=(puc or pont* univ* cat* or cat* univ* or pont* cat* univ*)) and CI=rio de janeiro
rice	OG=rice univ or SG=rice univ

Tabela G.1: *Strings* de consulta no formato WOS ISI por universidade (continuação)

Universidade	*String* de consulta
stanford	OG=stanford univ*
syracuse	OG=syracuse univ or SG=syracuse univ
texas-austin	OG=univ texas or SG=univ texas
ufmg	OG=(ufmg or univ* fed* minas gerais or fed* univ* minas gerais or minas gerais fed* univ*)
ufpe	OG=(ufpe or univ* fed* pernambuco or fed* univ* pernambuco or pernambuco fed* univ*) or SG=(ufpe or univ* fed* pernambuco or fed* univ* pernambuco or pernambuco fed* univ*)
ufrgs	OG=(univ* fed* rio grande do sul or fed* univ* rio grande do sul) or SG=(univ* fed* rio grande do sul or fed* univ* rio grande do sul)
ufrj	OG=(ufrj or univ* fed* rio de janeiro or univ* fed* rio janeiro or fed* univ* rio de janeiro or fed* univ* rio janeiro) or SG=(ufrj or univ* fed* rio de janeiro or univ* fed* rio janeiro or fed* univ* rio de janeiro or fed* univ* rio janeiro)
unb	OG=(unb or univ brasilia) or SG=(unb or univ brasilia)
unicamp	(OG=(CAMPINAS UNIV* OR UNIV* EST* CAMPINAS OR UNIV* CAMPINAS OR ST* UNIV* CAMPINAS OR UNICAMP) or SG=(CAMPINAS UNIV* OR UNIV* ESTADUAL CAMPINAS OR UNIV* CAMPINAS OR ST* UNIV* CAMPINAS OR UNICAMP))
usp	(OG=(st* univ* s* paulo or univ* s* paulo or s* paulo univ* or s* paulo st* univ*) or SG=(st* univ* s* paulo or univ* s* paulo or s* paulo univ* or s* paulo st* univ*)) and CU=(brazil or brasil)
waterloo	OG=univ waterloo
yale	OG=yale univ

Apêndice H

Exemplos de relações de coautoria na rede de coautoria extensa

Nesta seção são apresentados exemplos da metodologia de redução da ambiguidade na denominação dos autores 3.5.2 proposta para a ferramenta. Dois exemplos são situações em que a rede de coautoria melhora a recuperação de relações de coautoria melhorando a precisão dos dados obtidos na rede de coautoria inicial (Fase 3). Um terceiro exemplo demonstra que a rede de coautoria extensa pode melhorar também os resultados do mecanismo de agrupamento utilizado para construção da rede de autoria (Fase 2). Ao final há uma argumentação sobre as vantagens e desvantagens de proceder ao agrupamento por denominação para rede de autoria (Fase 2) separado por universidade.

H.1 Denominações em agrupamentos de autores pivôs

Na descrição da construção da rede de coautoria inicial fica claro que esta rede de coautoria prioriza estabelecer as relações de coautoria, mas pode desprezar algumas relações com poucos indícios de serem corretamente atribuídas.

Neste exemplo, partindo da fase 3 (rede de coautoria inicial), há a seguinte situação: o artigo *ISI:000180946400001* não foi considerado de autoria de *Cláudia Maria Bauzer Medeiros* (autora pivô). O artigo *ISI:000180946400001* possui a seguinte relação de denominações de autores: *M.J. Blin* e *C.B. Medeiros*.

O sistema ao realizar a construção da rede de coautoria inicial encontra poucos indícios de que *C.B. Medeiros* neste artigo correspondesse ao autor pivô *Cláudia Maria Bauzer Medeiros* membro do quadro de professores do Inst. de Computação da Unicamp. Ou seja, esta denominação de autor não possui especificado o email do autor pivô na lista de emails do artigo, e ao nome abreviado *C.B. Medeiros* não está associado uma denominação de nome completo. Os únicos indícios desta relação de autoria são a área de pesquisa em

comum e similaridade entre nomes abreviados. Ou seja, a área de pesquisa do Instituto de Computação da Unicamp é *CS* (*Computer Science*) e o conjunto de categorias de pesquisa do artigo denota que este corresponde a esta área de pesquisa: *Computer Science, Information Systems*. Quanto ao valor da similaridade entre as denominações, há apenas a similaridade baseada nas denominações de nome abreviado *C.B. Medeiros* e a abreviação correspondente a *Cláudia Maria Bauzer Medeiros*: *C. M. B. Medeiros* cujo valor é 94 em valores que variam de 0 (nenhuma similaridade) a 100 (similaridade máxima). Os critérios de mínimos indícios de similaridade exigem: email do autor pivô no artigo ou caso não haja o nome completo correspondente que a similaridade seja maior que 97.

Este exemplo mostra como *C. M. B. Medeiros* do artigo *ISI:000180946400001* pode ser atribuído à *Cláudia Maria Bauzer Medeiros* utilizando um caminho pela rede de coautoria. Primeiramente, a denominação *C. M. B. Medeiros* do artigo *ISI:000180946400001* será ponto de partida para o caminho na rede de coautoria partindo do agrupamento da autora pivô *Cláudia Maria Bauzer Medeiros* da Unicamp. O objetivo é atingido neste caso encontrando uma nova relação de coautoria entre dois autores pivôs. Neste caso, entre *Cláudia Maria Bauzer Medeiros* e outro coautor pivô, uma das possibilidades é através da autoria deste artigo *ISI:000180946400001*. Dentre os coautores apenas *J. Wainer* deste artigo foi atribuído a um autor pivô, no caso *Jacques Wainer* que é do quadro efetivo do Inst. de Computação da Unicamp. Felizmente a denominação *J. Wainer* casa perfeitamente com a de *J. Wainer* (a abreviação de *Jacques Wainer*). Esta denominação de autor também tem poucos indícios de similaridade ou seja, somente o nome abreviado e área de pesquisa, mas o suficiente para lhe atribuir a autoria: a sua similaridade é 100 (máxima) e o artigo é da mesma área de pesquisa da instituição ao qual *Jacques Wainer* é afiliado.

Sendo assim a denominação *J. Wainer* no agrupamento de denominações do autor pivô *Jacques Wainer* será o ponto de partida na busca de um caminho passando pelas relações de coautoria para chegar à outra denominação associada à autora pivô *Cláudia Maria Bauzer Medeiros*. As relações de coautoria não são relações estritas, conforme pode ser visto no algoritmo de construção da rede de coautoria inicial, ou seja, elas representam a possibilidade da relação de coautoria entre dois agrupamentos de denominações de autores. Em particular alguns destes agrupamentos estão associados com autores pivôs (lista de pesquisadores fornecidos ao sistema por universidade ao qual estão afiliados). Estas relações foram construídas tomando como base as similaridades entre as denominações dentro de cada agrupamento ao qual passou um caminho na rede de coautoria que partiu e chegou a agrupamentos de autores pivôs (veja mais detalhes na seção 3.7).

Estes requisitos mínimos são heurísticas para evitar a perda de precisão na recuperação da informação. A outra exigência estabelecida pelo algoritmo de construção da rede de coautoria extensa (Fase 4) é que o ponto de chegada de um caminho na rede de coautoria

seja uma denominação com um mínimo de indícios de similaridade e que já faça parte da rede de coautoria. Além disso que esta tenha pelo menos uma relação de coautoria com outro agrupamento de denominações de autor pivô. Neste exemplo, através do artigo *ISI:000231970300004*, há uma relação de coautoria entre *Cláudia Maria Bauzer Medeiros* e outro autor pivô *Edmundo Roberto Mauro Madeira*.

O artigo *ISI:000231970300004* tem dentre as denominações de autores *C.B. Medeiros* que não possui nome completo, mas o email de *Cláudia Maria Bauzer Medeiros* na Unicamp consta na lista de emails dos autores do artigo. Sendo assim é um excelente ponto de retorno ao agrupamento de denominações dela, no caso similaridade máxima (100).

A denominação *E. Madeira* que também está entre as denominações de autores do artigo *ISI:000231970300004* já foi atribuído pela rede de coautoria a *Edmundo Roberto Mauro Madeira* (um autor pivô) e o seu email consta na lista de emails deste artigo o que configura similaridade máxima (100).

Assim, através deste mesmo artigo a coautoria entre o *Edmundo Roberto Mauro Madeira* e *Cláudia Maria Bauzer Medeiros* tem possibilidade máxima 100.

Logo, há um caminho que sai de uma denominação de poucos indícios de ser atribuída a *Cláudia Maria Bauzer Medeiros*, e passando pelas possíveis relações de coautoria retorna a outra denominação de artigo neste agrupamento com indícios suficientes de autoria, desta forma resulta num valor de possibilidade ponderado de 100*100/100 = 100 para a relação de coautoria entre *Cláudia Maria Bauzer Medeiros* e *Jacques Wainer* e fica atribuída a *Cláudia Maria Bauzer Medeiros* autoria do artigo *ISI:000180946400001*.

H.2 Autores pivôs e agrupamentos que tem coautoria com autores pivôs

Este é o segundo exemplo de situações em que a rede de coautoria melhora a recuperação de relações de coautoria melhorando a precisão dos dados obtidos na rede de coautoria inicial (Fase 3).

Ao agrupamento de denominações de *Y. Kohayakawa* não foi atribuído a autoria do artigo *ISI:000223641900003* e este artigo figura na rede de coautoria por intermédio da alta similaridade e fortes indícios de que a denominação *FK Miyazawa* corresponda ao autor pivô *Flávio Keidi Miyazawa* (email, área de pesquisa) e os outros autores *P. Raghavan* e *Y. Wakabayashi* não têm correspondência com a lista de autores pivôs.

Dentre os candidatos de chegada o caminho bem sucedido mais com curto com grau de possibilidade suficiente foi através do artigo *ISI:000189502700015* publicado em coautoria por *F. K. Miyazawa* (com fortes indícios de ser de autoria do *Flávio Keidi Miyazawa*: email, área de pesquisa e similaridade de nomes 100) e também de autoria de *A. V.*

Moura com fortes indícios de ser de autoria de *Arnaldo Vieira Moura* (email, área de pesquisa, similaridade de nomes 100).

A tabela abaixo mostra a execução do algoritmo:

Tabela H.1: Segundo exemplo de relações de coautoria obtidas pela rede de coautoria extensa

Ids Ag.		Denominação de autores (autor pivô) (indícios de similaridade H.3)		
S	C	Saída	Artigo H.2	Chegada
0	1	F.K. Miyazawa (Flávio Keidi Miyazawa) (SEA 100)	ISI:...003	Y. Kohayakawa
1	2	Y. Kohayakawa (100)	ISI:...005	E. Laber (Eduardo Sany Laber) (SEA 100)
2	3	E.S. Laber (Eduardo Sany Laber) (SEA 100)	ISI:...023	C. Bornstein (Cláudio Thomas Bornstein) (SA 91)
3	4	C. Bornstein (Cláudio Thomas Bornstein) (SEA 100)	ISI:...008	N. Maculan
4	5	N. Maculan (100)	ISI:...004	C.C. de Souza (Cid Carvalho De Souza) (SEA 100)
5	6	C.C. de Souza (Cid Carvalho De Souza)(SA 89)	ISI:...029	A.V. Moura (Arnaldo Vieira Moura) (SA 100)
6	0	A.V. Moura (Arnaldo Vieira Moura) (SEA 100)	ISI:...015	F.K. Miyazawa (Flávio Keidi Miyazawa) (SEA 100)

Tabela H.2: Identificadores dos Artigos do segundo exemplo

Abreviado	Completo
ISI:...003	ISI:000223641900003
ISI:...005	ISI:000221969900005
ISI:...023	ISI:000236886100023
ISI:...008	ISI:000228123800008
ISI:...004	ISI:000236740000004
ISI:...029	ISI:000229801300029
ISI:...015	ISI:000189502700015

Na Tabela H.1 em sua primeira linha a denominação de autor *F.K. Miyazawa* que é parte do agrupamento do autor pivô *Flavio Keidi Miyazawa* com indícios suficientes de similaridade e é um dos autores do artigo *ISI:000223641900003*, onde um dos coautores é a denominação *Y. Kohayakawa*. A rede de coautoria inicial não identificou que o agrupamento de denominações a qual *Y. Kohayakawa* faz parte seja autor deste artigo. O objetivo deste exemplo é mostrar que há uma possibilidade de coautoria entre o autor pivô *Flávio Keidi Miyazawa* e o agrupamento de denominações a qual faz parte *Y. Kohayakawa*.

No agrupamento de denominações de *Y. Kohayakawa*, existe outra instância *Y. Kohayakawa* com similaridade máxima (100) com a anterior e que corresponde ao artigo *ISI:000221969900005*.

Tabela H.3: Legendas dos indícios de similaridade entre autores do segundo exemplo

Abreviado	Completo
Valor Numérico	Similaridade de nome, onde 0 é mínima e 100 é máxima
E	Email do autor pivô presente na lista de emails do artigo
S	Possui denominação de autor abreviada
A	Área de Pesquisa do artigo corresponde à área de pesquisa do autor pivô

Neste artigo, uma das denominações de autor é *E. Laber* com fortes indícios de corresponder ao autor pivô *Eduardo Sany Laber* com similaridade máxima (100). Esta relação de coautoria entre o agrupamento de *Y. Kohayakawa* e um autor pivô é condição mínima que seja considerado início da busca na rede de coautoria para retornar a este agrupamento através do agrupamento pivô de *Flávio Keidi Miyazawa*.

No agrupamento de denominações de *Eduardo Sany Laber* há outra *E.S. Laber* também com similaridade máxima com este autor pivô. Esta última denominação corresponde ao artigo *ISI:000236886100023* que tem como um dos coautores *Cláudio Thomas Bornstein* por intermédio da denominação *C. Bornstein* que possui similaridade 91 com este autor pivô. No agrupamento de denominações do autor pivô *Cláudio Thomas Bornstein* há outra denominação *C. Bornstein* com fortes indícios de similaridade (100) com este autor pivô e foi encontrada no artigo *ISI:000228123800008* onde uma das denominações de autores é *N. Maculan*.

No agrupamento de denominações de *N. Maculan* há outra com similaridade máxima com esta (100) advinda do artigo *ISI:000236740000004*. Neste artigo, um dos coautores é a denominação *C.C. de Souza* com fortes indícios de ser do autor pivô *Cid Carvalho De Souza* com similaridade máxima (100). No agrupamento de denominações do autor pivô *Cid Carvalho De Souza* há outra denominação de autor correspondente ao artigo *ISI:000229801300029*, mas que não possuindo o email teve como base apenas a similaridade de nomes (*C.C.de Souza* com C.C.D. Souza) cujo valor é 89. Outro autor deste artigo é a denominação *A.V. Moura* que tem fortes indícios de ser do autor pivô *Arnaldo Vieira Moura* com similaridade máxima (100).

Assim, o algoritmo fecha o circuito, pois neste agrupamento de denominações está um dos pontos de chegada *A.V. Moura* do artigo *ISI:000189502700015* com forte indícios de ser de autoria deste autor pivô e este artigo possui como denominação de coautor *F.K. Miyazawa* com fortes indícios de ser de autoria do autor pivô *Flavio Keidi Miyazawa* conforme era o intuito da busca pela rede de coautoria. O valor ponderado para caminho entre as relações de possibilidade de autoria é de $90\frac{89}{100} = 80$, desprezando os fatores com

valor 100 que não alteram o resultado.

Desta forma é atribuída a possibilidade de coautoria 80 entre o autor pivô *Flávio Keidi Miyazawa* e o autor *Y. Kohayakawa*, e a autoria do artigo *ISI:000223641900003* passa também a ser atribuída ao autor *Y. Kohayakawa*.

H.3 Denominações de autores pivôs que ficaram fora do agrupamento

Este terceiro exemplo demonstra que a rede de coautoria extensa pode melhorar também os resultados do mecanismo de agrupamento utilizado para construção da rede de autoria (Fase 2).

Ao agrupamento de denominações de *André Carlos Ponce De Leon Ferreira De Carvalho* não foi atribuída à denominação correspondente ao artigo *ISI:000220901200001*. Este artigo figura na rede de coautoria por intermédio da alta similaridade e fortes indícios da denominação de autor *A.C.B. Delbem* corresponder a *Alexandre Cláudio Botazzo Delbem*, ou seja, têm a mesma área de pesquisa e similaridade de nomes com o valor máximo (100). Os outros autores são: *A. De Carvalho* e *N. G. Bretas* todos sem correspondentes na lista de autores pivôs.

Apesar do artigo *ISI:000220901200001* possuir na sua lista de emails dos autores o email de *André Carlos Ponce De Leon Ferreira De Carvalho*, dentre as denominação de autores a mais similar *A. De Carvalho* possui uma similaridade muito baixa 76 e portanto não consta no agrupamento do autor pivô *André Carlos Ponce De Leon Ferreira De Carvalho*. Este fato indica a necessidade de verificar se há ao menos indícios obtidos via rede de coautoria para que este artigo possa ser a ele atribuído.

Alexandre Claudio Botazzo Delbem publicou outros artigos com *André Carlos Ponce De Leon Ferreira De Carvalho*, como por exemplo, o artigo *ISI:000221714200051* onde consta a denominação de coautor *A.C.P.L.F. Carvalho* com fortes indícios de ser de autoria de *André Carlos Ponce De Leon Ferreira De Carvalho*, ou seja, email, área de pesquisa e similaridade de nomes máxima (100). Sendo assim, o algoritmo pode atribuir a autoria do artigo *ISI:000220901200001* ao *André Carlos Ponce De Leon Ferreira De Carvalho*. Infelizmente não foram feitos muitos testes com esta abordagem, pois os critérios para adota-la precisam ser melhorados, pois esta situação descrita é muito rara e seria melhor aplicável depois das iterações da fase 4 não mais agregarem informações de autoria ou de coautoria.

H.4 Processamento separado por universidade dos agrupamentos de autores

Este exemplo será utilizado em uma argumentação sobre o ponto favorável de não realizar o agrupamento de denominações de autores apenas baseado em dados de artigos obtidos de cada universidade. O ponto favorável da alternativa oposta é claramente a possibilidade de processamento ser realizado em paralelo por universidade. No entanto, isto pode resultar em perda de importantes relações de coautoria, como será apresentado neste exemplo, ou o custo de encontrá-las através de um mecanismo de unificação é pior do que processar tudo em conjunto. Por outro lado, numa situação em que o volume de dados exija, pode ser adotada esta alternativa apesar do tempo de processamento ser maior ou assumindo que a perda de informações como aceitável.

Neste exemplo é retomado o artigo *ISI:000189239900004* do primeiro exemplo e este artigo é identificado pelas consultas da extração da base de dados bibliográfica como produção científica da Unicamp apenas, as outras instituições são "não identificadas" pelo sistema. Por outro lado, o artigo *ISI:000237081600060* foi identificado pelas consultas como produção científica da Northwestern Univ. apenas, as outras instituições são "não identificadas" pelo sistema. Estes artigos têm em comum o pesquisador cuja denominação de autor é *Voisard, A* e na forma que foi concebido sistema ele pertence a uma instituição "não identificada". Apesar de ele estabelecer coautoria com *Cláudia Maria Bauzer Medeiros* da Unicamp e com *Goce Trajcevski* e *Peter Scheuermann* da Univ. NorthWestern esta unificação que permite o caminho na rede de coautoria *Cláudia Maria Bauzer Medeiros* → *A. Voisard* → *Goce Trajcevski* só pode ser identificada se o agrupamento contiver as denominações correspondentes a artigos que estão em produção científica de universidades diferentes (Northwestern Univ. e Unicamp), caso contrário o sistema teria custo muito grande para calcular a possibilidade de unificá-los. Este custo é mais alto (ou igual) ao que já de início considera as denominações de artigos de todas as universidades para realizar o agrupamento. Os agrupamentos de autores pivôs têm um processo simples de unificação, basta que cada identificador numérico associado a cada denominação de autor seja único independente da universidade que está processando. Como os autores pivôs têm um identificador único, eles servem de referência para unificar agrupamentos de mesmo autor pivô em processamento de agrupamento feito em paralelo. A situação difícil ocorre, por exemplo, no caso do *A.Voisard* que tem identificação *id1* devido ao artigo *ISI:000189239900004* que só consta nos dados da Unicamp e *id2* no artigo *ISI:000237081600060* que só consta nos dados da Univ. NorthWestern. Quando os dados destas duas universidades são processados em conjunto o agrupamento { *id1,id2* } é encontrado por similaridade exata de nomes, e foi observando manualmente nos dados dos dois artigos consta o email de *A.Voisard*, mas não é o caso quando o processamento

é feito em separado como será descrito a seguir.

O autor pivô *Goce Trajcevski* professor do quadro de professores da área de computação na Univ. North-Western que publicou também o artigo *ISI:000238525500002* com os coautores *H. Cao* e *O. Wolfson*. No entanto, *H. Cao* não consta na lista do quadro de professores da Univ. Illinois e *Ouri Wolfoson* é professor na Univ. Illinois segundo o parâmetro informado ao sistema. Este artigo também consta nos dados obtidos na consulta da Univ. NorthWestern por constar na lista de afiliações a instituições deste artigo um nome de universidade que casa com esta consulta. Como já foi dito que *Goce Trajcevski* casou por similaridade com a denominação *G. Trajcevski* constante na lista de autores deste artigo.

Quando o processamento dos dados de cada universidade é feito em paralelo, haverá para cada autor pivô da base um agrupamento diferente em cada resultado do agrupamento por universidade Por exemplo: *Goce Trajcevski* com *id3* nos agrupamento encontrados na Univ. NorthWestern, devido ao artigo *ISI:000237081600060* constante nesta base, então a denominação de autor *G .Trajcevski* será associada a ele com *id4* e por conseguinte obterá um agrupamento que contenha pelo menos {*id3,id4* }. Nos dados da Unicamp nenhuma denominação é similar a este autor pivô então haverá um agrupamento {*id3*} (contendo apenas a identificação do autor pivô). Por outro lado, na Univ. Illinois o artigo *ISI:000238525500002* que tem a denominação de autor *G .Trajcevski* com um *id5* (novamente um id único na base), então produzirá um agrupamento que contenha pelo menos {*id3,id5* }

A Unificação os dados da Univ. Illinois com a NorthWestern, tomando como referência o autor pivô *Goce Trajcevski* que tem *id3* ficaria da seguinte forma: { *id3,id4,* ...} ∪ {*id3,id5,* ...} = { *id3,id4,id5,* ...}.

No caso de *A. Voisard* haverá o agrupamento {*id1*} nos agrupamento calculados com os dados da Unicamp e agrupamento { *id2,* ...} obtidos nos dados da Univ. NorthWestern. Sendo assim, não há uma referência comum para unificá-los, inviabilizando a atribuição de autoria *ISI:000189239900004* e *ISI:000237081600060* a um "autor" definido pelo agrupamento de denominações { *id1,id2* }. Isto configura uma desvantagem que deve ser considerada quando é necessário adotar o processamento em separado dos dados de cada universidade.

Referências Bibliográficas

[1] igraph. http://cneurocvs.rmki.kfki.hu/igraph/, 2007-2009. 3.4

[2] Python. http://www.python.org/, 2007-2009. 3.4

[3] Scipy. http://www.scipy.org/, 2007-2009. 3.4

[4] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, Addison-Wesley, 1999. 2.6

[5] I. Bhattacharya and L. Getoor. Iterative record linkage for cleaning and integration. In *Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*, pages 11–18, Paris, France, June 2004. ACM. 2.1, 2.5

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, 1998. 2.1, 2.3.2

[7] Elsevier B.V. Portal da scopus. http://www.scopus.com/scopus/home.url, 01/2008. 1

[8] Stella Chen, Dmitri V. Kalashnikov, and Sharad Mehrotra. Adaptive graphical approach to entity resolution. In *ACM/IEEE Joint Conference on Digital Libraries*, 2007. 2.2, 2.3

[9] Zhaoqi Chen, Dmitri V. Kalashnikov, and Sharad Mehrotra. Exploiting relationships for object consolidation. In *Proceedings of Information Quality in Informational Systems (ACM IQIS)*, pages 47–58, June 2005. 2.5, 2.5.1

[10] Aaron Clauset. Finding community structure in very large networks. *Physical Review*, 2005. 2.7.2

[11] CNPq. Portal da plataforma lattes. http://lattes.cnpq.br, 01/2008. 4.6

[12] ShanghaiRanking Consultancy. Academic ranking of world universities. www.arwu.org. em 08/2010. 1, 2.4.1, 2.8, 4.4, 5, D, E

[13] Rodrigo De Castro and Jerrold Grossman. Famous trails to paul erdös. *The Mathematical Intelligencer*, 21(3):51—-53, 1999. 2.3

[14] Wouter de Nooy, Andrej Mrvar, and Vladimir Batagelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2005. 2.1, 2.2

[15] Henrique Przibisczki de Oliveira. Ranking de publicações baseado na extração de textos da internet. Master's thesis, IC da UNICAMP, 12 2009. 2.6.2, 3.6.3, 4.6, 5

[16] J. Diesner and Kathleen M. Carley. Using network text analysis to detect the organizational structure of covert networks. In *Proceedings of the North American Association for Computational Social and Organizational Sciences (NAACSOS), Conference*, June 2007. 7, 8

[17] Xing Dong, Alon Halevy, and Jayant Madhavan. Reference reconciliation in complex information spaces. *ACM IQIS*, June 2005. 2.5

[18] TSL Education. World university ranking 2009. http://www.timeshighereducation.co.uk/, em 08/2010. 2.8, 4.6

[19] SCImago Research Group. Portal do scimago journal & country rank. http://www.scimagojr.com/, 08/2010. 2.1, 2.8, 4.6

[20] H. Han, C. L. Giles, and H. Zha. A model-based k-means algorithm for name disambiguation. In *Proceedings of the Second International Semantic Web Conference (ISWC)*, Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, 2003. 2.1

[21] H. Han, H. Zha, and C. L. Giles. Name disambiguation in author citations using a k-way spectral clustering method. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 334–343. ACM, 2005. 2.1

[22] Hui Han, Lee Giles, Hongyuan Zha, Cheng Li, and Kostas Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 296–305, New York, NY, USA, 2004. ACM. 2.1, 3.6.2

[23] D. Kerridge. The interpretation of rank correlations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):257–258, 1975. 4.6

[24] Stefan Klink, Patrick Reuther, Alexander Weber, Bernd Walter, and Michael Ley. Analysing social networks within bibliographical data. In S. Bressan, J. Küng, and R. Wagner, editors, *Database and Expert Systems Applications (DEXA)*, volume 4080 of *LNCS*, pages 234–243, 2006. 2.1, 2.5

[25] Alberto H. F. Laender, Carlos J. P. de Lucena, José Carlos Maldonado, Edmundo de Souza e Silva, and Nivio Ziviani. Assessing the research and education quality of the top brazilian computer science graduate programs. *SIGCSE Bull.*, 40(2):135–145, 2008. 4.6

[26] Julia Lane. Let's make science metrics more scientific. *Nature*, 464, March 2010. 4.6

[27] Dongwon Lee, Jaewoo Kang, Prasenjit Mitra, C. Lee Giles, and Byung-Won On. Are your citations clean? new scenarios and challenges in maintaining digital libraries. *Communication of the ACM (CACM)*, 2007. 1

[28] Mong Li Lee, Wynne Hsu, and Vijay Kothari. Cleaning the spurious links in data. *IEEE Intelligent Systems*, 19(2):28–33, 2004. 2.1, 2.5

[29] Xiaoming Liu, Johan Bollen, Michael L. Nelson, and Herbert Van de Sompel. Co-authorship networks in the digital library research community. *Information Processing & Management*, 41(6):1462–1480, 2005. Special Issue on Infometrics. 2.3.1

[30] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *J. Mach. Learn. Res.*, 2:419–444, 2002. 2.6.2

[31] U.S.News & World Report LP. Usnews rankings. http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-computer-science-schools/rankings, em 08/2010. 1, 2.8, 5

[32] Bradley Malin. Unsupervised name disambiguation via social network similarity. In *Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security, in conjunction with the SIAM International Conference on Data Mining*, pages 93–102, Newport Beach, CA,USA, 2005. 2.5

[33] Jesús Pascual Mena-Chalco and Roberto Marcondes Cesar Junior. Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, 15:31–39, 2009. 4.6

[34] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64(1):016131, Jun 2001. 2.3

[35] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1):016132, Jun 2001. 2.3

[36] Jean W. A. Oliveira, Alberto H. F. Laender, and Marcos André Gonçalves. Remoção de ambiguidades na identificação de autoria de objetos bibliográficos. Master's thesis, IC da UFMG, 2005. 2.6.1

[37] ByungWon On, Dongwon Lee, Jaewoo Kang, and Prasenjit Mitra. Comparative study of name disambiguation problem using a scalable blockingbased framework. In *Proceedings of the 5th International Conference on Digital Libraries (ACM/IEEE-CS) joint conference on Digital libraries table of contents, Tools & techniques*, identifying names of people and places, pages 344–353, 2005. 1

[38] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *Physics Society*, December 2005. 2.7.2, 2.7.2, 2.7.4

[39] Thomson Reuters. Portal do web of science. http://portal.isiknowledge.com/portal.cgi, 01/2008. 1, 2.1, 2.4, 5

[40] Thomson Reuters. Thomson journal list. http://scientific.thomson.com/mjl/, em 01/2008. 2.4.1

[41] P. Sprent and N. C. Smeeton. *Applied Nonparametric Statistical Methods*, chapter 7. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC, third edition, 2001. 1, 2.8.1

[42] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining.* Addison Wesley. 2.6, 2.7.3, 2.7.4, 3.6.2, 4.1

[43] Paul D. Valz, A. Ian McLeod, and Mary E. Thompson. Cumulant generating function and tail probability approximations for kendall's score with tied rankings. *The Annals of Statistics*, 23(1):144–160, February 1995. 2.8.1

[44] Jr. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963. 2.7

[45] Stanley Wasserman and Katherine Faust. *Social Network Analysis.* Cambridge University Press, 1994. 2.1, 2.3.2, 3.5.2, 5

[46] D. J. Watts. *Six Degrees: The Science of a Connected Age.* W. W. Norton & Company, New York, USA, 2004. 2.1

[47] Hattendorf Westney and C. Lynn. Historical rankings of science and technology: A citationist perspective. *The Journal of the Association of History and Computing*, 1(1):1, June 1998. 2.4

[48] Rui Xu and D. II Wunsch. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, may 2005. 2.7.1, 2.7.3

[49] Yiming Yang. An evaluation of statistical approaches to text categorization. *Inf. Retr.*, 1(1-2):69–90, 1999. 2.7.5, 2.7.5

[50] Ying Zhao and George Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, June 2004. 2.7.4